# Depth Image Resolution Enhancement using Discrete Wavelet Transform and Convolution Neural Networks

Seyed Mehrdad Mahdavi[1]*, Mohsen Ashourian[2]
1- Department of Computer Engineering, Najafabad Branch, Islamic Azad University, Najafabad, Iran.
Email: Mehrdad5087@gmail.com (Corresponding author)
2- Department of Electrical Engineering, Majlesi Branch, Islamic Azad University, Majlesi, Iran.
Email: ashourian@iaumajlesi.ac.ir

Received: September 2020 Revised: November 2020 Accepted: December 2020

**ABSTRACT:**
The depth image plays an increasingly important role in fundamental research and daily applications, with the reducing the price and increasing the number of affordable and portable depth cameras. Infrared sensors or depth sensors are widely used to control dynamic and static 3D scenes. However, the depth image quality is limited to low-quality images, as the infrared sensor does not have high resolution. Therefore, given the problems and the importance of using 3-D images, the quality of these images should be improved in order to provide accurate images from depth cameras. In this paper a resolution enhancement method of depth images using convolutional neural networks is considered. A convolutional neural network with a depth of 20 and three layers and a pre-trained neural network is used. We developed the system and tested its performance for two datasets, Middlebury and EURECOM Kinect Face. Results show for EURECOM Kinect Face images, PSNR improvement is approximately 7 to 16 dB and for Middlebury images the PSNR improvement is about 6 to 12 dB.

**KEYWORDS:** Depth Camera Images, Image Enhancement, Super Resolution, Convolution Neural Networks.

## 1. INTRODUCTION

Depth enables a user to rapidly create detailed 3D reconstructions of an indoor scene. The Kinect depth camera consist of a standard color camera and an infrared camera. The color camera captures a color image, as the name implies, from the environment and scene, which is used in subsequent processing to improve the depth of field images, and the depth camera estimates the depth by reflecting infrared light. In these cameras, a light source transmits infrared light with a dot pattern, then the sensors, which are the heart of the camera, take a recursive pattern and estimate the distance based on the length of sweet time of the light [1].

Three-dimensional information from a scene includes position information (depth information) and texture information. While texture information is easily captured by color cameras, it is not easy to obtain depth information. In addition, the obtained depth information requires pre-processing to be used for subsequent processing, (Improvements), since, the depth maps recorded by the depth cameras have very low resolution compared to the color image. These depth maps also suffer from various damages such as low sampling, loss of structure along the depth discontinuities, and accidental loss in smooth areas, which make these images noisy with lack of sharpness and ultimately their quality would be reduced. Such destructions have hampered their practical application. Depth cameras have errors when shooting objects with special features such as sharp edges, and their error increases in very bright environments. These problems are caused by changes in ambient brightness, scene geometry, ambient heat, and elevated sensor temperatures over time, so given the current problems and the wide usage of infrared sensors to control application and games and obtain information from dynamic and 3-D scenes, the image quality of these cameras has to be improved. Despite the widespread use of these images, their quality is limited to low-quality and low-resolution images, because the infrared sensor does not have high resolution and the images produced by it have noise. Therefore, due to the existing problems and the importance of using depth images, the quality of these images should be improved in order to provide accurate images using depth cameras [2].

Usually in systems like Kinect, having a depth image and a color image, the depth image has lower resolution

Paper type: Research paper
DOI: https://doi.org/10.29252/mjtd.10.1.9
How to cite this paper: S. M. Mahdavi, M. Ashourian, **“Depth Image Resolution Enhancement using Discrete Wavelet Transform and Convolution Neural Networks”**, *Majlesi Journal of Telecommunication Devices*, Vol. 10, No. 1, pp. 9-15, 2021.

compare to the color image. It seems possible to find solution to reconstruct the lower resolution depth image to a higher resolution one using the color information based on intelligent signal processing methods. In this paper at first, we follow a wavelet-based approach to up-sample and interpolate the depth image. Later we use a convolution neural network to further

Section 2 provides a review on literature on super resolution methods and convolution neural networks. Section 3 explains the proposed method. Section 4 shows the result of simulation and finally Section 5 concludes the paper.

## 2. LITERATURE REVIEW

The spatial resolution is a key parameter for any digital imaging system, and it refers to the pixel density in an image. The image spatial resolution is first limited by the imaging sensors or the imaging acquisition system. Also, the hardware cost of a sensor increases with the increase of sensor density or related image pixel density.

Another approach for enhancing the resolution is by employing various signal processing tools. These techniques are specifically referred to as Super-Resolution (SR) reconstruction. There was a significant speared in this field [3]. Approaches using Frequency Domain, Bayesian, and Interpolation [4] have been applied to SR techniques. Image resolution enhancement in the wavelet domain is a relatively new research topic and recently many new algorithms have been proposed [5]–[6]. Discrete wavelet transforms (DWT) [7] is one of the recent wavelet transforms used in image processing. DWT decomposes an image into different subband images, namely low-low (LL), low-high (LH), high-low (HL), and high-high (HH).

We further import the interpolated image to a convolution neural network (CNN). There are three main reasons for using these types of networks:

- Convolutional neural networks with deep architecture have the capability and flexibility to describe image properties [8].
- Another notable advantage of these types of networks is their learning methods, which include Rectifier Linear Unit (ReLU) [8], batch normalization, and residual learning [9]. In these papers, this type of learning is introduced for classification and recognition tasks, but can also be used as a future research area to reduce noise and speed up the learning process of the network.
- Convolutional neural networks using parallel computing are well compatible with modern, powerful GPUs, which can be used to reduce their running time.

The most important reason for using CNN to reduce image distortion is that it does not require to estimate original image, and the distortion is estimated directly. This is done by the difference between the distorted image and the clear one. It should be noted that this paper uses color information-based estimators because if only depth sensors are used, the depth results depend on how they are navigated and for these results to be accurate and produce high quality images, very good navigation should be performed on them. To overcome these problems, a color camera that produces a high-quality color image can be used to improve the quality of the low-quality depth image produced by the depth sensor. In fact, a color image is used to take advantage of adjacent of the dots in the color image and the associated depth image, increasing their ability to measure local similarities, turning them from piecewise into patches and having less processing complexity than other methods.

## 3. RESEARCH METHODOLOGY

Our system has two steps, we follow a similar method like what proposed in [9] for image resolution improvement and then give the output of that system to a deep neural network proposed which is explained later. Fig. 1 shows an overview of the developed system. Fig. 2 shows the up-sampling method proposed in [7] and we adapted in our developed system. The proposed technique uses DWT to decompose a low-resolution depth image into different subbands. Then the three-high frequency subbands have been interpolated using bicubic interpolation. The high frequency subbands obtained by SWT of the input image are being incremented into the interpolated high frequency subbands in order to correct the estimated coefficients. In parallel, the input depth image is also interpolated separately. Finally, corrected interpolated high frequency subbands and interpolated input image are combined by using inverse DWT (IDWT) to achieve a high-resolution output image.

In the second step the reconstructed depth image is imported to a convolution neural network. In terms of network architecture design, our proposed method is a modified VGG network [10]. VGG can be used to reduce image noise. In the proposed method the depth of network is adjusted based on patch sizes. In terms of model learning, the residual learning formulation has also been selected and combined with patch normalization to accelerate training and improve noise reduction performance. Similar to the method used in Reference [10], in this paper, the size of the convolutional filters is assumed $3 \times 3$, except that all pooling layers are removed. In the convolutional neural network architecture, the observed noisy image of $y = x + v$ is the input the network where x is the original image and υ is the additive noise. The image can have any dimension, or it can even be gray or colored. Unlike other articles that estimate the clear and noise-free image, we estimate noise that is assumed to be of an unknown nature. Then, with the estimated noise, we can easily obtain a clear, noise-free image through $x = y - R(y)$. For this purpose, the mean square error criterion is used.

$$\text{PSNR}= 20\log\frac{(x-y)}{255} \qquad (1)$$

Fig. 3 shows the proposed neural network architecture for learning.

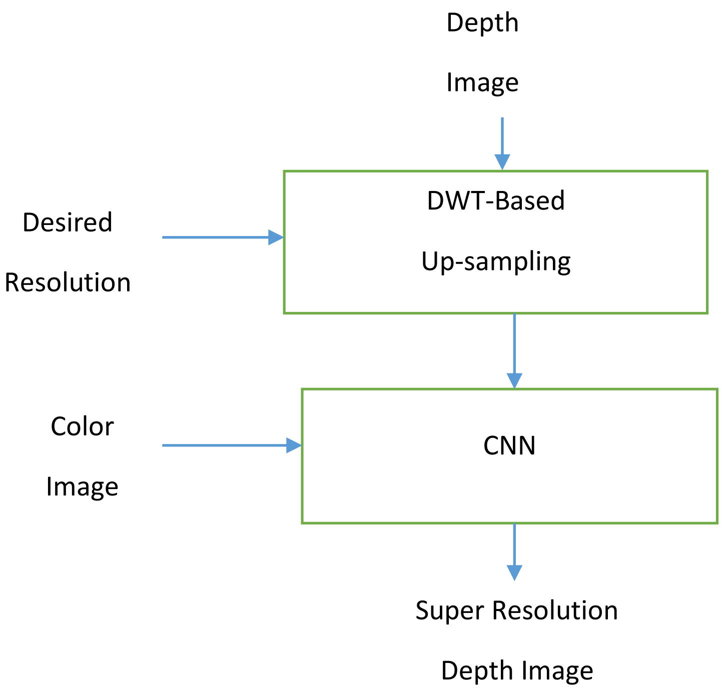


**Fig. 1.** The Basic Block Diagram of the proposed System.

.

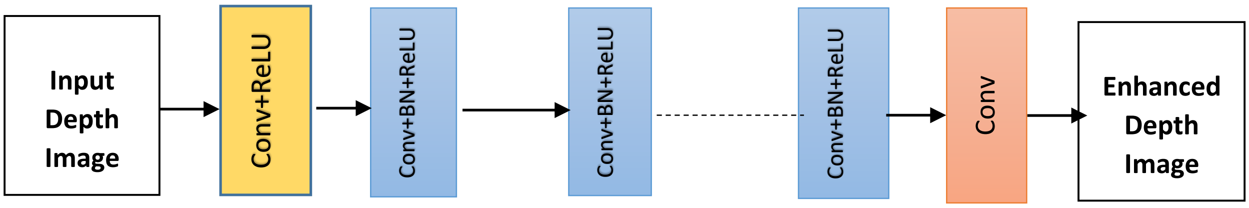


**Fig. 2.** Proposed convolutional neural network architecture.

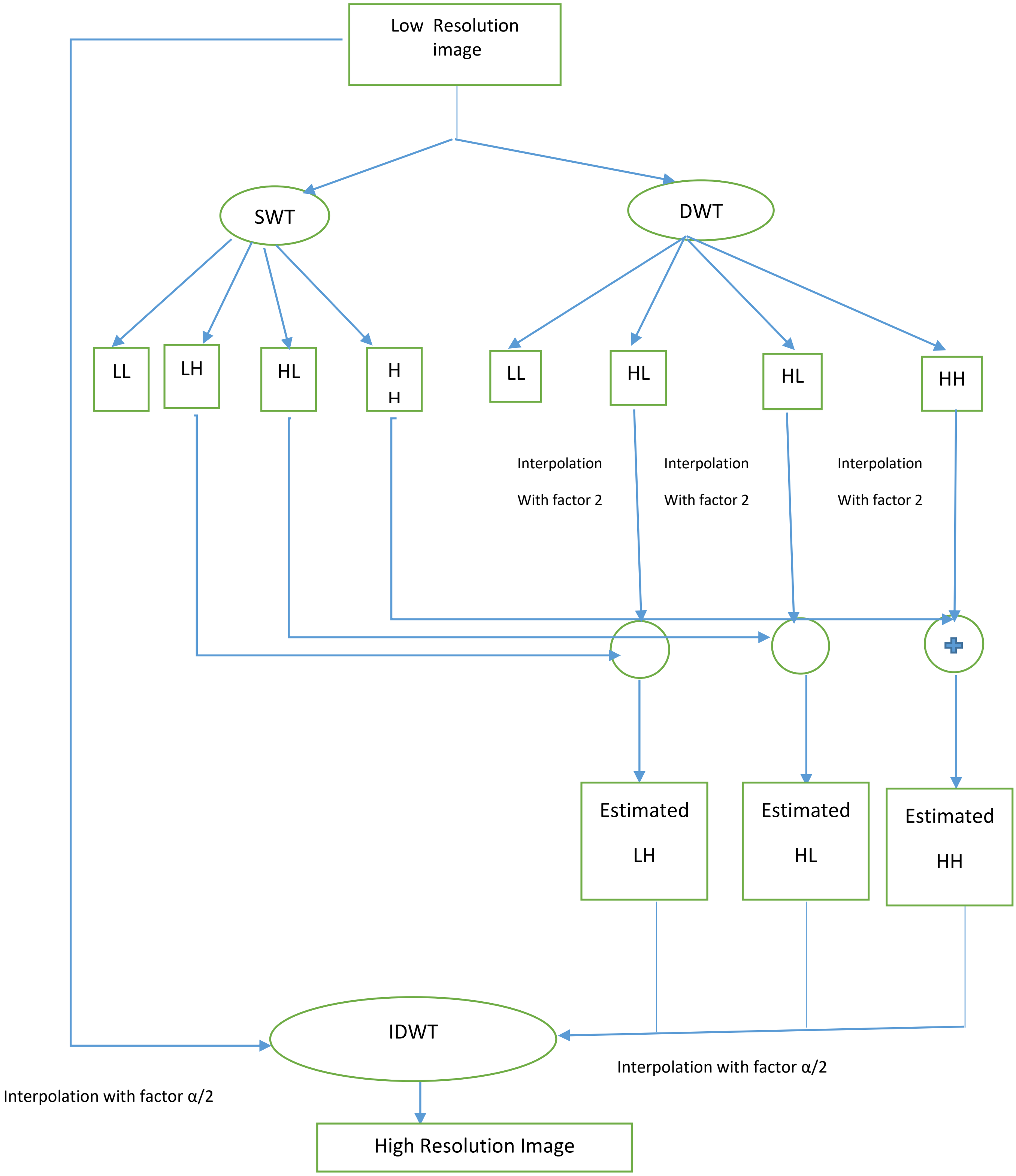
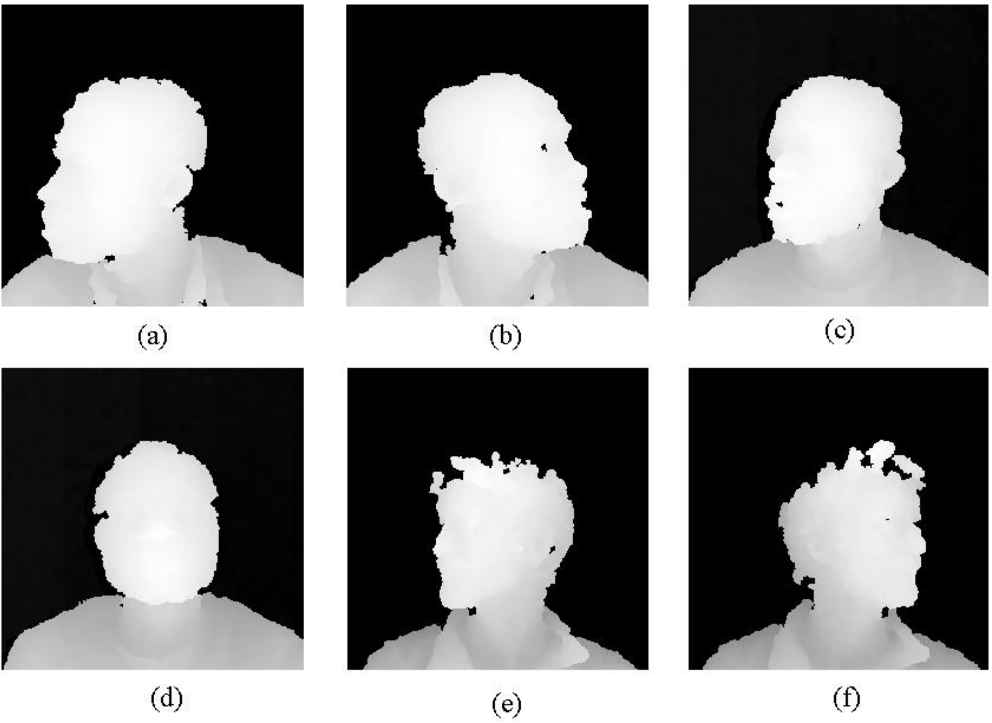

Low Resolution image
SWT
DWT
LL
LH
HL
H H
LL
HL
HL
HH
Interpolation With factor 2
Interpolation With factor 2
Interpolation With factor 2
Estimated LH
Estimated HL
Estimated HH
IDWT
Interpolation with factor α/2
Interpolation with factor α/2
High Resolution Image

**Fig. 3.** DWT-Based Up-sampling System (adapted from [7].



m

**Fig. 4.** Six Image samples of the EURECOM Kinect Face dataset.

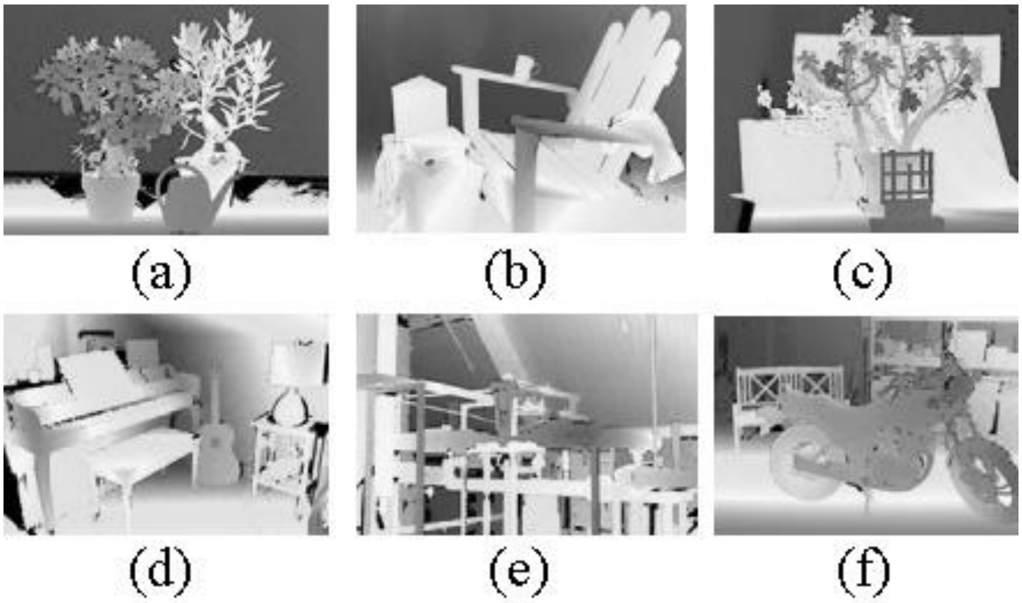


**Fig. 5.** Six image samples of the Middlebury dataset.

As can be seen from Fig. 2, the up-sampled image from the left enters the system, and on the right, the residual image is obtained. It is clear from the figure that there are three layers in the convolutional neural network that are shown in different colors. These three layers are as follows:

1. Conv + ReLU: This layer is the first layer of convolutional neural network shown in yellow. This layer contains 64 filters with a size of $3\times3\times c$, whose task is to create 64 feature mappings; and there are also 64 rectifier linear units (ReLU) that provide nonlinearity. The

parameter $C$ represents the image type or number of channels of the image, so if $c=1$, the image is gray and if $c=3$, the image is colored.

2. Conv + BN + ReLU: These layers are in the second layer to the $D-1$ layer and are shown in blue. In this layer, 64 filters with a size of $3\times3\times64$ are used, as well as patch normalization introduced in Reference [12] between convolution and ReLU.
3. Conv: This is the last layer in orange (Fig. 2). It has a number of $C$ filters with a size of $3\times3\times64$ used to form the output.

In many machine vision applications, the output and the input image need to be equal in size. Some references in the final stage consider a process to reduce the size of the image and make it as the size of input image [18]. In the proposed method, before the image reaches the final layer or the convolution layer, it is ensured by using a condition that the output image and the input image are equal and the work of equalizing them is done in the convolution layer.

## 4. RESULTS AND DISCUSSION

The proposed method has been applied to different depth images and their resolution improved in one set of experiment from 128×128 to 256×256 pixel, and in another test from 64×64 to 256×256. For this purpose, two types of datasets are considered, which are introduced below:

1. EURECOM Kinect Face dataset: This dataset contains the faces of different people who were photographed using a Kinect camera. In this dataset, 52 people are photographed at 9 different angles and their depth and color image are collected. [19]. An example of images of the EURECOM Kinect Face dataset is shown in Fig. 3.
2. 2-The second dataset is known as Middlebury, in which the collection contains images of various sights and scenes. This dataset was compiled in 2001, 2003, 2005, 2006 and 2014 [20-23]. An example of Middlebury dataset is shown in Fig. 4.

In our simulation, a pre-trained convolutional neural network with 500 images and the learning method mentioned in the reference [24] are used for training, all images have dimensions of 256×256.

In this case, the size of the used patches is assumed to be 50*50, so that the total number of 128*3000 patches were used to train the noise removal model. The number of network depths is 20 and the loss function is (1) selected to learn the network. The weights of the network were calculated by the method used in the reference [25] and the gradient descent method with a weight delay of 0.0001. In addition, network training over 50 iPOCs has been able to build noise removal models.

The simulations are performed in MATLAB software. After running, a convolutional network with 33 layers is created which each layer's information is stored in a structure. This information includes layer type, layer weights, weight delay, learning rate, and so on.

Middlebury dataset results: Six images of Middlebury dataset is selected which are shown in Fig. 5.

EURECOM Kinect Face Datasheet Results: For this datasheet, six images were selected and manually down-sampled. Later the down-sampled images enters the developed system as input.

In Table 1 the result of the PSNR calculations of the difference between original depth image and reconstructed depth image obtained from the network is calculated and provided.

**Table 1**. Results of PSNR Calculations of Original Input Depth Image reconstructed depth Image dataset after reducing its resolution by 2 and then up sampling by 2.

| σ | EURECOM Kinect Face Dataset | Middlebury Dataset |
|---|---|---|
| | PSNR | PSNR |
| Image 1 | 35.2 | 37.0 |
| Image 2 | 34.2 | 36.8 |
| Image 3 | 34.8 | 35.7 |
| Image 4 | 32.6 | 34.5 |
| Image 5 | 33.5 | 32.1 |
| Image 6 | 37.0 | 32.8 |

**Table 2**. Results of PSNR Calculations of Original Input Depth Image reconstructed depth Image dataset after reducing its resolution by 4 and then up sampling by 4.

| σ | EURECOM Kinect Face Dataset | Middlebury Dataset |
|---|---|---|
| | PSNR | PSNR |
| Image 1 | 32.6 | 31.3 |
| Image 2 | 30.0 | 30.6 |
| Image 3 | 29.7 | 29.9 |
| Image 4 | 29.5 | 31.9 |
| Image 5 | 30.1 | 28.5 |
| Image 6 | 31.2 | 28.8 |

## 5. CONCLUSION

In this paper, a method for depth images up-sampling is presented using DWT and convolutional neural networks. The proposed model is applied to two Middlebury and EURECOM Kinect Face datasets with reduced resolution and the up-sampled images are compared with original depth images. The results show the high similarity of reconstructed depth image to the original one.

## REFERENCES

[1] S. Foix, G. Alenya, and C. Torras, "**Lock-in time-of-flight (ToF) cameras: A survey,**" *IEEE Sensors Journal,* Vol. 11, No. 9, pp. 1917-1926, 2011.

[2] E Eichhardt, Ivan, Dmitry Chetverikov, and Zsolt Janko. "**Image-guided ToF depth upsampling: a survey**." *Machine Vision and Applications* 28.3-4, pp. 267-282, 2017.

[3] Park, Sung Cheol, Min Kyu Park, and Moon Gi Kang. "**Super-resolution image reconstruction: a technical overview**." *IEEE signal processing magazine* 20.3, pp. 21-36, 2003.

[4] W. K. Carey, D. B. Chuang, and S. S. Hemami, "**Regularity-preserving image interpolation**," *IEEE Trans. Image Process.*, Vol. 8, No. 9, pp. 1295–1297, Sep. 1999.

[5] Y. Piao, I. Shin, and H. W. Park, "**Image resolution enhancement using inter-subband correlation in wavelet domain**," in *Proc. Int. Conf. Image Process.*, Vol. 1, pp. I-445–448, 2007.

[6] Temizel and T. Vlachos, "**Image resolution upscaling in the wavelet domain using directional cycle spinning**," J. Electron. Imag., Vol. 14, No. 4, 2005.

[7] Demirel, Hasan, and Gholamreza Anbarjafari. "**Image resolution enhancement by using discrete and stationary wavelet decomposition**." IEEE transactions on image processing 20.5, pp. 1458-1460, 2010.

[8] Yao, Guangle, Tao Lei, and Jiandan Zhong. "**A review of Convolutional-Neural-Network-based action recognition.**" *Pattern Recognition Letters* 118, pp. 14-22, 2019.

[9] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "**Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising**," *IEEE Transactions on Image Processing,* Vol. 26, No. 7, pp. 3142-3155, 2017.

[10] Krizhevsky, I. Sutskever, and G. E. Hinton, "**Imagenet classification with deep convolutional neural networks**," in *Advances in neural information processing systems,* pp. 1097-1105, 2012.

[11] Min, Rui, Neslihan Kose, and Jean-Luc Dugelay. "Kinectfacedb: **A kinect database for face recognition**." *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 44.11, pp. 1534-1548, 2014.

[12] Baker, S., Scharstein, D., Lewis, J., Roth, S., Black, M., Szeliski, R.: **A database and evaluation methodology for optical flow**. IJCV 92, pp. 1–31, 2011.