IMAGE GUIDED DEPTH ENHANCEMENT VIA DEEP FUSION AND LOCAL LINEAR REGULARIZATION

Jiang Zhu  Jing Zhang  Yang Cao  Zengfu Wang

Department of Automation, University of Science and Technology of China
{zjl30129, zjwinner}@mail.ustc.edu.cn, {forrest, zfwang}@ustc.edu.cn

ABSTRACT

Depth maps captured by RGB-D cameras are often noisy and incomplete at edge regions. Most existing methods assume that there is a co-occurrence of edges in depth map and its corresponding color image, and improve the quality of depth map guided by the color image. However, when the color image is noisy or richly detailed, the high frequency artifacts will be introduced into depth map. In this paper, we propose a deep residual network based on deep fusion and local linear regularization for guided depth enhancement. The presented scheme can effectively extract the correlation between depth map and color image in the deep feature space. To reduce the difficulty of training, a specific layer of network which introduces a local linear regularization constraint on the output depth is designed. Experiments on various applications, including depth denoising, super-resolution and inpainting, demonstrate the effectiveness and reliability of our proposed approach.

Index Terms—depth enhancement, deep residual network, local linear regularization, deep feature space

1. INTRODUCTION

High-quality depth maps are very important in computer vision applications, such as 3DTV, 3D reconstruction, hand pose estimation and robot navigation. With the development of consumer depth cameras such as Kinect and Xtion Pro, it is more convenient to capture the depth map of scene. However, depth maps captured by these devices are often noisy and incomplete at edge regions due to the reflection and absorption of structure light or the viewpoint disparity between multi-sensors.

Since depth cameras often provide a pair of color and depth (RGBD) image of a scene, various methods [1] [2] [3] [4] [5] [6] have been proposed to enhance the quality of depth under the guidance of color image. These methods are based on the assumption that there is a co-occurrence of edges in depth map and its corresponding color image [7] [8]. However, this assumption does not always hold in practice. When the color image is noisy or richly detailed, the high frequency component of color image will be introduced into depth map. These generated artifacts will seriously affect the quality of depth map.

Many works introduce deep learning into various image processing applications, such as image denoising [9], image super-resolution [10] [11] [12] [13], and image restoration [14]. Their successes benefit from underlying-feature representative ability of deep neural networks. Very recently, Zhang et al. [15] proposed a deep CNN based method which demonstrates the feasibility of an end-to-end approach to depth enhancement. Their method uses max pooling to widen receptive field. However, the output depth map is blurry. For depth enhancement applications, such as denoising and super-resolution, the sharp edges of depth map are very important.

In this paper, we propose a CNN-based framework for guided depth enhancement. Our goal is to learn the underlying correlation between depth maps and color images, and then to use the correlation to enhance the quality of depth map. To achieve this goal, we need to overcome the following challenges. 1) To preserve the sharp edges in depth map, no pooling can be used in the proposed network. An alternative solution to improve the performance of deep CNN is to increase the network depth by adding new layers. However, this will introduce more parameters and increase the risk of overfitting. 3) Depth map and color image have different noise level and data distribution. It will not work well for directly using depth map and color image as the input to jointly train the network. A feature level fusion strategy is required for the learning process.

To address the above challenges, we propose a deep residual network for guided depth enhancement. The proposed deep network consists of three components: depth branch, intensity branch and deep feature fusion part. Fig.1 illustrates the details of our proposed network. Our proposed framework has the following advantages:

1) We present a deep residual network for image guided depth enhancement, which can well extract the underlying correlation between depth map and color image in the deep feature space.

2) A deep fusion strategy is adopted. The low-level features for depth and color image are firstly extracted in each
branch, respectively. And then the features are combined together and fused into high-level features. This scheme effectively suppresses the influence of different noises and data distributions in depth map and color image.

3) To avoid overfitting, a specific layer which introduces a local linear regularization constraint on the output depth is presented. It significantly improves convergence speed and accuracy.

2. PROPOSED METHOD

2.1. Proposed architecture

We propose a deep residual network which employs a single residual unit to learn the residual image, as shown in Fig.1. We refer to the proposed deep residual network for guided depth enhancement as DRECNN. Our network consists of three components: depth branch, intensity branch and deep feature fusion part. Moreover, to achieve a wide receptive field, we set the number of convolution layers to 20 for all of the depth enhancement applications.

The Conv, ReLU, BN and SR represent convolution layers, rectified linear units, batch normalization layer and a specific regularized layer, respectively. There are three types of layers in our network. (1) Conv+ReLU: using in the first layer of depth branch (D) and intensity branch (Y). (2) Conv+BN+ReLU: using from the second layer to the nineteenth layer. (3) Conv+SR: using in the last layer to reconstruct the output depth map. The specific regularized layer (SR) will be introduced in detail in Sec. 2.3.

2.2. A deeply fusion strategy for joint-feature extraction

Depth map and color image have different noise level and data distribution. Directly using depth map and intensity image as the input of network will bring great interference to training process. To overcome this problem, we propose a deep fusion strategy for joint-feature extraction.

In our proposed network, the first eight layers extract the low-level features of the depth maps and the intensity images, respectively. Then the next eleven layers combine the two types of low-level features together and fuse into the high-level joint-features. This deep fusion strategy can efficiently extract the correlation between depth map and color image even in the presence of high-level noises.

2.3. A local linear regularization constraint

The image guided filter [8] starts from a local linear model as follows:

$$q_i = a_k I_i + b_k, \forall i \in \omega_k.$$  (1)

Here $I$ is a guidance image and $q$ is the filter output. The $\omega_k$ is the filter window and $(a_k, b_k)$ are the linear coefficients which are assumed to be constant in the window $\omega_k$. The filter output $q$ is a linear transform of the guidance image $I$.

Inspired by the local linear model, we propose a specific regularization (SR) layer which is shown as Fig.2. Specifically, we first insert a convolution layer to output the linear coefficient map $a$ and $b$, and then compute $aY + b$ as the results of
linear model. Finally, the residual map $aY + b - D$ is computed and supervised by the ground truth label (We use $aY + b + D$ instead in our network. Indeed, it is an equivalent expression as $aY + b - D$, since the network will learn the opposite linear coefficients adaptively.). The SR layer is exploited in the output part of our proposed network, which serves as a constraint on the output depth map. This method can effectively boost the performance of depth enhancement and reduce the risk of overfitting.

2.4. Loss function definition

For the learning process of our proposed networks, we use a Euclidean-based distance function as the loss function:

$$L(\Theta) = \frac{1}{2N} \sum_{i=1}^{N} \| F(D_i, Y_i; \Theta) - (D_i - D_i^{GT}) \|^2_F.$$  \hspace{1cm} (2)

We denote $F$ as the mapping function and $\Theta$ as our network parameters, where $D_i$ is the input depth map, $Y_i$ is the input intensity image, $D_i^{GT}$ is the ground truth depth map, and $N$ represents the amount of the training samples.

3. APPLICATIONS AND EXPERIMENTAL RESULTS

In this part, we conducted a series of experiments on various applications including depth denoising, super-resolution and inpainting to demonstrate the effectiveness and reliability of our proposed DRECNN.

3.1. Data preprocessing and augmentation

Training and testing data: We used the same training dataset for all of the depth enhancement applications. We chose 40 RGBD images from MPI Sintel depth dataset [16] and 21 RGBD images from Middlebury dataset [17][18][19]. In the training images, 55 images were used for training and the remaining 6 images were used for validation. Then we chose the rest images of Middlebury dataset which are not used in the training as the test images. Note that we used bilateral filter [20] to pre-process the groundtruth depth maps of Middlebury dataset since it still had some missing values.

Data augmentation: For data augmentation, we rotated the original training images with 90°, 180°, 270° and also flipped them upside-down. In this way, we obtained 8× more samples.

3.2. Experimental setting

Parameter setting: We set the same parameter for different applications. The size of the patches was 50×50. We used Adam algorithm to train our DRECNN. The momentum was 0.9, a mini-batch size was 128 and the weight decay was 0.0005. We set the learning rate from $1e^{-3}$ to $1e^{-6}$ which was gradually decayed for 1.8e + 5 iterations. The kernel size of convolution layers was 3×3 and we padded zeros for convolution layers to keep the same size as the input depth.

Network training: Our DRECNN was implemented using caffe [21]. We used a Tesla M40 GPU for the training and testing stage. To reduce training time and memory burden, we converted the input color images from RGB to YCbCr and only used the luminance component (Y) for training.

3.3. Performance analysis on regularized layer

To demonstrate the significance of the specific regularized (S-R) layer, we performed two contrasting experiments. One experiment used the regularized layer to add constraint on the output depth, the other one used the normal output layer. Fig.3 presented the two training results. As can be seen, the regularized layer significantly improved the convergence speed. Both training loss and test loss were much lower compared with using normal output layer.

3.4. Experimental results on various applications

A. Depth Blind Denosing

For depth blind denosing, we added different levels of additive white Gaussian noise(AWGN) to the depth map pathes of the training dataset, where the noise level $\sigma$ was set as 10, 15, 20, 25 and 30. We also added the additive white Gaussian noise(AWGN) to intensity image (Y) with standard variance of 15. The denoising comparisons with the state-of-the-art methods in terms of PSNR are summarized in table 1 (The best results are highlighted in bold font ). And the visual inspection results are shown in Fig.4.

As can be seen in the Table 1, our proposed network produced superior denoising results than the state-of-the-art methods such as BM3D [22] and NLGBT [23] in terms of PSNR. As shown in Fig.4, our proposed method reduced artifacts effectively while preserving more details. Moreover, our DRECNN achieved better performance in presence of high-level noises.

B. Depth Super-resolution

Fig. 3. The training and test loss of two different methods.
Table 1. Denoising comparison in terms of PSNR (dB)

<table>
<thead>
<tr>
<th>Images</th>
<th>Methods</th>
<th>σ=10</th>
<th>σ=15</th>
<th>σ=20</th>
<th>σ=25</th>
<th>σ=30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cones</td>
<td>BM3D [22]</td>
<td>40.56</td>
<td>37.49</td>
<td>35.28</td>
<td>33.81</td>
<td>32.75</td>
</tr>
<tr>
<td></td>
<td>NNLGBT [23]</td>
<td>42.84</td>
<td>39.18</td>
<td>36.53</td>
<td>34.43</td>
<td>32.97</td>
</tr>
<tr>
<td></td>
<td>DRECNN</td>
<td>43.15</td>
<td>40.53</td>
<td>39.02</td>
<td>37.46</td>
<td>36.02</td>
</tr>
<tr>
<td>Teddy</td>
<td>BM3D [22]</td>
<td>41.36</td>
<td>38.33</td>
<td>36.12</td>
<td>34.45</td>
<td>33.25</td>
</tr>
<tr>
<td></td>
<td>NNLGBT [23]</td>
<td>42.29</td>
<td>39.38</td>
<td>36.71</td>
<td>34.62</td>
<td>33.42</td>
</tr>
<tr>
<td></td>
<td>DRECNN</td>
<td>42.84</td>
<td>41.22</td>
<td>39.64</td>
<td>38.31</td>
<td>36.90</td>
</tr>
<tr>
<td>Sawtooth</td>
<td>BM3D [22]</td>
<td>46.04</td>
<td>43.51</td>
<td>41.84</td>
<td>40.16</td>
<td>39.13</td>
</tr>
<tr>
<td></td>
<td>NNLGBT [23]</td>
<td>48.41</td>
<td>45.30</td>
<td>43.22</td>
<td>41.71</td>
<td>40.01</td>
</tr>
<tr>
<td></td>
<td>DRECNN</td>
<td>49.05</td>
<td>46.28</td>
<td>45.14</td>
<td>43.21</td>
<td>42.06</td>
</tr>
</tbody>
</table>

Fig. 4. Denoising results of the Cones with noise level 10.

For depth map super-resolution, the high-resolution depth maps were down-sampled and up-sampled with the scaling factors 4 and 8 using the method of Mandal et al. [24]. Additive white Gaussian noise (AWGN) with standard variance of 5 was added to the low-resolution depth maps. We also added the additive white Gaussian noise (AWGN) to intensity image (Y) with standard variance of 15.

The comparison results with the state-of-the-art methods in terms of the root mean squared errors (RMSE) are summarized in Table 2, where we refer to Mandal et al.’s method [24] and Aodha et al.’s method [25] to PS and EB. The visual inspection results are shown in Fig.5. As can be seen, our proposed method achieved better performance with the larger scaling factors. Moreover, our method suppressed the noise effectively while preserving the sharp edge.

Table 2. Depth super-resolution comparison in terms of RMSE for σ = 5

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Aloe</td>
<td>8.09</td>
<td>5.73</td>
<td>5.09</td>
<td>13.05</td>
<td>9.05</td>
<td>7.59</td>
</tr>
<tr>
<td>Baby</td>
<td>5.06</td>
<td>3.78</td>
<td>3.05</td>
<td>8.43</td>
<td>5.64</td>
<td>4.83</td>
</tr>
<tr>
<td>Cones</td>
<td>6.11</td>
<td>4.49</td>
<td>4.23</td>
<td>10.00</td>
<td>6.94</td>
<td>5.67</td>
</tr>
<tr>
<td>Plastic</td>
<td>4.47</td>
<td>3.19</td>
<td>2.42</td>
<td>8.32</td>
<td>4.65</td>
<td>3.25</td>
</tr>
<tr>
<td>Teddy</td>
<td>5.04</td>
<td>3.77</td>
<td>3.50</td>
<td>8.38</td>
<td>5.50</td>
<td>4.32</td>
</tr>
<tr>
<td>Venus</td>
<td>2.66</td>
<td>2.63</td>
<td>2.18</td>
<td>4.35</td>
<td>3.43</td>
<td>2.96</td>
</tr>
</tbody>
</table>

C. Depth Inpainting

For depth inpainting, we used the method of LRMC [26] to pre-process the depth maps as training data. Moreover, we added additive white Gaussian noise (AWGN) with standard variance of 25 to the intensity image (Y). The comparison results with the state-of-the-art methods in term of the PSNR are summarized in Table 3 (The best results are highlighted in bold font). The visual inspection results are shown in Fig.6. In most cases, our DRECNN achieved higher PSNR than the state-of-the-art methods. As can be seen in Fig.6, DE-CNN [15] tends to generate blurry edges because of max pooling in their network. Our DRECNN produced sharper edges and preserved more details than the other methods, e.g. LRMC [26] and Shen et al.’s method [27].

Table 3. Depth inpainting comparison in terms of PSNR (dB)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Art</th>
<th>Dolls</th>
<th>Reindeer</th>
<th>Laundry</th>
<th>Baby</th>
<th>Wood</th>
<th>Teddy</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE-CNN [15]</td>
<td>33.84</td>
<td>40.67</td>
<td>35.46</td>
<td>38.11</td>
<td>40.28</td>
<td>40.84</td>
<td>39.77</td>
</tr>
<tr>
<td>LRMC [26]</td>
<td>33.10</td>
<td>41.86</td>
<td>36.13</td>
<td>38.59</td>
<td>39.00</td>
<td>39.96</td>
<td>40.77</td>
</tr>
<tr>
<td>Shen et al. [27]</td>
<td>33.42</td>
<td>42.89</td>
<td>37.19</td>
<td>38.71</td>
<td>39.32</td>
<td>41.54</td>
<td>40.89</td>
</tr>
<tr>
<td>DRECNN</td>
<td>34.02</td>
<td>41.59</td>
<td>38.18</td>
<td>39.17</td>
<td>41.13</td>
<td>41.74</td>
<td>41.21</td>
</tr>
</tbody>
</table>

Fig. 5. Depth map super-resolution results of the Teddy with upscaling factor 8.

Fig. 6. Depth map inpainting results of the Baby.

4. CONCLUSION

In this paper, we proposed a deep residual network-based framework for image guided depth enhancement, in which a deep feature fusion strategy and a specific regularization layer are adopted. The deep feature fusion strategy could suppress the influence of different noises distributions in depth map and color image. And the local linear regularization constraint on the output depth could improve the training speed and precision. The experimental results on various applications, including depth denoising, super-resolution and inpainting, demonstrated that our network could produce favorable depth enhancement results even with high-level noises. Acknowledgments. This work is supported by the Natural Science Foundation of China(61472380).
5. REFERENCES


