Intrinsic Image Decomposition using Deep Convolutional Network

Hyeongseok Son
POSTECH
sonhs@postech.ac.kr

Seungyong Lee
POSTECH
leesy@postech.ac.kr

Abstract

This paper proposes a deep convolutional network for intrinsic image decomposition from a single image. To maintain the visual details of a result, we approach the problem from the perspective of filtering, without pooling for downsampling and unpooling for recovering the original resolution. We use several 1D convolutional layers with alternating horizontal and vertical directions to effectively capture the large structures of a scene. We also add regularization terms to the loss functions to reduce overfitting, and use synthetically rendered images to overcome the lack of training data. Experiments show that our approach achieves visually pleasing separation of shading and reflectance.

1. Introduction

Intrinsic image decomposition addresses the problem of decomposing an image into reflectance and shading, which is inherently ill-posed. It has been studied extensively due to the applications in computer vision and graphics. Intrinsic image decomposition of a single image [5] is typically solved by optimization based on Retinex theory [7]. To improve the results, recent methods exploit additional information, such as depth and surface normal [2, 3].

Although deep learning has become popular in image processing, it is not straightforward to apply deep learning to intrinsic image decomposition. It would need a network model to consider both global and local structures of a scene, and a big dataset to cover a variety of scenes.

A common approach to apply deep learning to image processing, e.g, semantic segmentation, is to extract high-level features by downsampling (pooling) and then upsample (unpooling) the features to the original resolution. This approach is effective for obtaining overall structural information, but is not appropriate for intrinsic image decomposition due to possible loss of details.

In this paper, we propose a novel solution for intrinsic image decomposition of a single image using a deep convolutional network. To preserve the details in the decomposition results, we use a filtering network without pooling and unpooling, preventing the image resolution changes in the network. To consider the global structure of a scene, as well as local information, we use several 1D convolutional layers with alternating horizontal and vertical directions, providing a large receptive field in the filtering network. For training our network, we use a synthetic dataset where ground truth reflectance and shading layers are given with high quality rendered images. In addition, we add the regularization terms from the Retinex model to the loss function to reduce overfitting in training the network.

2. Proposed Method

Network model Our network consists of convolution layers and ReLU layers without any pooling layers. To increase the size of the receptive field effectively, we use 10 convolution layers with alternating 1D kernels in horizontal (1×41) and vertical (41×1) directions. The size of the resulting receptive field is about 200×200.

An image I is decomposed into reflectance R and shading S, satisfying

$$I = R * S. \tag{1}$$

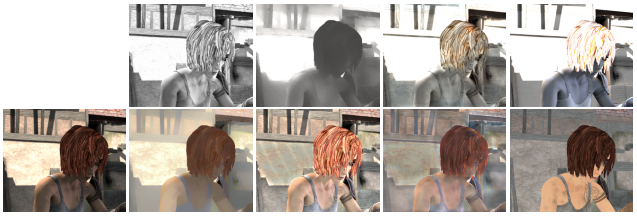
In the logarithmic domain, reflectance R can be represented as a *residual* of I when we have computed shading S.

$$log(R) = log(I) - log(S) \tag{2}$$

We use this residual structure in our network design, and our network has the shading output layer and then the reflectance output layer is defined by the residual.

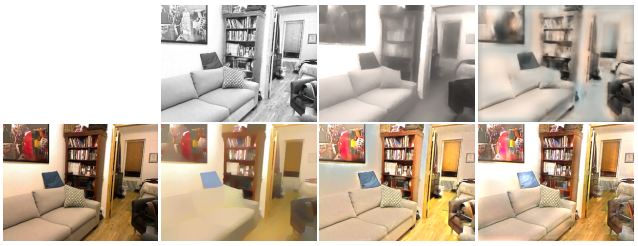
Our network has two loss layers for shading and reflectance. We add a regularization term of the Retinex model to the Euclidean loss function of each layer. Shading should be smoothed by L_2 norm of the gradients, and reflectance should be piecewise smooth depending on chroma values. This addition of regularization terms to the loss functions helps the network be trained while trying to avoid overfitting under the limited amount of training data.

Dataset Generating a ground-truth dataset of real images for intrinsic image decomposition is a laborious task. With



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

Figure 1: Results on MPI SINTEL dataset. (a) input image (b) *Shen et al.* [5], (c) *Chen et al.* [2], (d) our result, (e) ground-truth.



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

Figure 2: Results on NYU v2 dataset. (a) input image, (b) *Shen et al.* [5], (c) *Chen et al.* [2], (d) our result.

help of the authors, we obtained reflectance and shading information of various rendered scenes (1014 images of size 1024×436 for 22 scenes) from MPI SINTEL dataset [1].

Training We tested different settings for training our network: reflectance only, shading only, and both reflectance and shading. When we trained the network with the ground-truth reflectance only, where shading was automatically computed by Eq. 2, the results were unsuccessful. It restores the rough structure of the reflectance layer but with almost no color information. Training with only the ground-truth shading produces reasonable results but with shading details blurred. We obtained the best results when we use both ground-truth reflectance and shading for training.

We use adjustable gradient clipping [4] to accelerate the training. The technique enables the training with a high learning rate by preventing gradient exploding.

3. Results

We implemented our method and tested with various images on Intel Core i7 CPU and NVIDIA Titan X GPU. Figs. 1 and 2 show results on MPI SINTEL and NYU v2 dataset [6], respectively.

Our network spends 0.3s for processing a 1024×436 image. Our result is slightly better in preserving the details of reflectance than a local optimization method [5]. The method using additional depth information [2] shows better handling of textures, but overall shading is similar to our result, as shown in Figs. 1 and 2. We also tested with a real example of an outdoor scene, as shown in Fig. 3.

(a) (b) (c)

Figure 3: Results on a real image. (a) input image, (b) *Shen et al.* [5], (c) our result.

4. Conclusion

This paper proposed a novel filtering-based network for dealing with intrinsic image decomposition. It generates better results than previous methods that use only local information, as the large receptive field in our network could reflect some global context. However, experimental results show that even the large receptive field may not suffice for complete handling of global context yet. In addition, our network cannot distinguish dark shadows and dark objects, and texture handling is another remaining problem.

Acknowledgements This work was supported by Institute for Information & communications Technology Promotion (IITP) grant (R0126-16-1078) and the National Research Foundation of Korea (NRF) grant (NRF-2014R1A2A1A11052779) both funded by the Korea government (MSIP).

References

[1] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. *In Proc. ECCV*, 2012.

[2] Q. Chen and V. Koltun. A Simple Model for Intrinsic Image Decomposition with Depth Cues. *In Proc. ICCV*, 2013.

[3] J. Jeon, S. Cho, X. Tong, and S. Lee. Intrinsic image decomposition using structure-texture separation and surface normals. *In Proc. ECCV*, 2014.

[4] J. Kim, J. K. Lee, and K. M. Lee. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. *In Proc. CVPR*, 2016.

[5] J. Shen, X. Yang, Y. Jiang, and X. Li. Intrinsic images using optimization. *In Proc. CVPR*, 2011.

[6] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from RGBD images. *In Proc. ECCV*, 2012.

[7] Q. Zhao, P. Tan, Q. Dai, L. Shen, E. Wu, and S. Lin. A closed-form solution to retinex with nonlocal texture constraints. *IEEE TPAMI*, 34(7):1437–1444, 2012.