Segmentation of Illuminated Areas of Light Using CNN and Large-Scale RGB+D Dataset for Augmented and Mixed Reality Systems[*]

Maksim Sorokin[0000–0001–9093–1690], Dmitriy Zhdanov[0000-0001-7346-8155] and Andrey Zhdanov[0000-0002-2569-1982]

ITMO University, St. Petersburg, Russia

vergotten@gmail.com, ddzhdanov@mail.ru, andrew.gtx@gmail.com

Abstract. This work is devoted to the problem of restoring realistic rendering for augmented and mixed reality systems. Finding the light sources and restoring the correct distribution of scene brightness is one of the key parameters that allows to solve the problem of correct interaction between the virtual and real worlds. With the advent of such datasets as, "LARGE-SCALE RGB + D," it became possible to train neural networks to recognize the depth map of images, which is a key requirement for working with the environment in real time. Additionally, in this work, convolutional neural networks were trained on the synthesized dataset with realistic lighting. The results of the proposed methods are presented, the accuracy of restoring the position of the light sources is estimated, and the visual difference between the image of the scene with the original light sources and the same scene. The speed allows it to be used in real-time AR/VR systems.

Keywords: Augmented Reality, Mixed Reality, Fully Convolutional Network, Segmentation, Deep Learning, CNN, Computer Vision Algorithms.

1 Introduction

Augmented and mixed reality systems are being used in many tasks, however the incorrect illumination of the virtual world objects may cause discomfort in the perception of the reality, in which objects of the real and virtual worlds are mixed and as a result this limits the time that a person can be in the mixed reality, and further it restricts the practical use of the mixed reality systems in various areas, for example - in education or training.

This article is devoted to convolutional neural network methods (CNN) for solving the global scientific problem in the field of the physically correct and effective restoration of illumination conditions and optical properties of real-world objects during the synthesis of images of the virtual world.

[*] This work was supported by the Russian Science Foundation (project No. 18 -79-10190).

First, this work is focused on determining the real power of illumination of the light flux and its position in an environment. For this a manually synthesized sample of images with realistic optical parameters of the medium were used. Although the sample consists of only 260 images (221 were used for training, and39 for the test), the neural network at the output classifies with good accuracy the real optical parameters of the illumination of the medium. These are usually divided by the strength of illumination into 5 classes, where the first is 0 lumens, which means it is not lit at all, while grade 5 is the source of illumination of an ordinary room lamp.

Moreover, for the reconstruction of the depth of an environment the "Large-Scale RGB + D Dataset" was used, which was obtained using Kinect v2 and Zed stereo camera and their disparity maps.

2 Related works

The idea of environment segmentation is already trivial, but every year new approaches, neural network architectures and datasets appear that improve the previous results. An analysis of outdoor lighting using a fully convolution network is presented in [1], [2], where analyzes panoramic images of the environment in an open air as input and embeds the image under these environmental conditions. The work [3] also analyzes the environment and builds the shadows of objects as they should be. The convolution network is also used in [4], but to determine where the object is located: outdoors or indoors. The following article [5] presents its own architecture and solves three different problems: predicting depth, evaluating surface normals, and semantic marking. Many works [6, 7, 8, 9] were aimed at detecting objects using convolutional neural networks.

Works [11,12,13,14] as well as this work are aimed at restoring lighting for augmented reality systems, but for different purposes and tasks. Among these articles, methods for analyzing direct lighting are considered, without taking into account the secondary lighting for objects of virtual reality. Also, these works are aimed at finding the direct observer in the augmented reality and lighting system relative to the user. The difference between the neural network described in this paper is that the data set is generated using a powerful renderer - “Lumicept” [10], which were used to train the neural network, restoring segmented sections of light similar to reference images with ground truth using the categorical cross entropy object function. The main task of the current work is to determine and classify the real illumination power of a real room and the light source position.

The “LARGE-SCALE RGB+D DATABASE” dataset contains synchronized RGBD frames from both Kinect v2 and Zed stereo camera. For the outdoor scene, they first generated disparity maps using an accurate stereo matching method and converted them using calibration parameters. A per-pixel confidence map of disparity is also provided. The scenes are captured at various places, e.g., offices, rooms, dormitory, exhibition center, street, road etc., from Yonsei University and Ewha University. This dataset has been used to train convolutional neural networks in projects [15] and in papers[16],

[17], [18], [19] "High quality 2D-to-multiview contents generation from large-scale RGB-D database".

3 Implementation

To train a neural network for classification depth maps a sorted dataset "Large-Scale RGB + D Dataset" of indoor images was used, which consists of 1609 images for training and 503 images for testing. These images were obtained using Kinect v2 and Zed stereo cameras, with the calculation of their disparity maps and restoration of the pixel confidence. Sorted "Large-Scale RGB + D Dataset" comes with high-resolution and low-resolution quality, in current work was used the low-resolution pack with squeezing all images to 224*224 pixels, so that it would be possible to train them on the architecture of VGG 16-Net.

The working distance of the Kinect is up to 6 meters. The level of gradation between the distances of measurement levels is about ~30 cm. In total, in this approach with the neural network of the VGG 16-Net architecture 20 output neurons were used, to produce distance measurement with a 20-level heat map. As a convolution network architecture, it was decided to use the VGG16 Net architecture because it was successfully used in many classification tasks, it consists of 5 blocks with convolution, pooling, and "ReLU" activation function between layers and the optimization method is "Nesterov".

For lighting classification method was used manually synthesized dataset of lighting of different rooms. The task of a fully convolutional network is to classify each pixel of an image into one class. That is, passing through all convolutional layers, the network characterizes a certain area of the image to one class in accordance with the power of illumination.

The architecture is presented on figure 1.

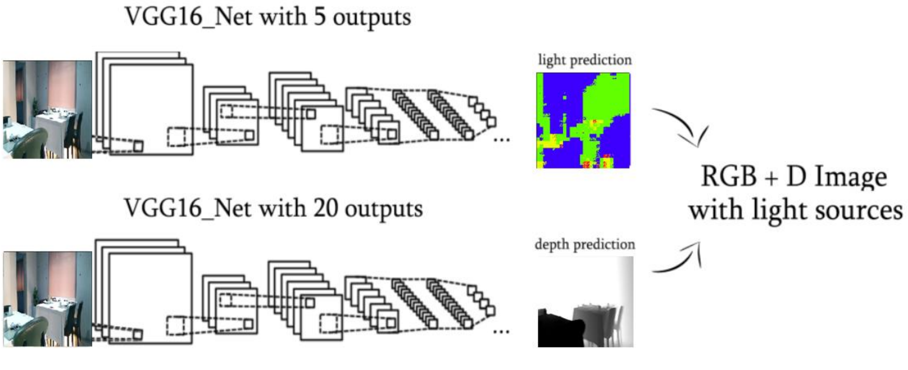

Fig.1. Were trained two VGG16-Net architectures with 5 and 20 neuron outputs, 5 output neurons for light classification and 20 – for depth. After classification, the both images unite and form one RGB+D image with corresponding light and depth.

The neural network training for the classification of light took 50 epochs, for depth map - 200 epochs. The training took place on a GeForse GTX 1080Ti video card. The original image was fed with 5 light area masks and 20 depth data masks. Dataset images are shown in figures 2 and 3.

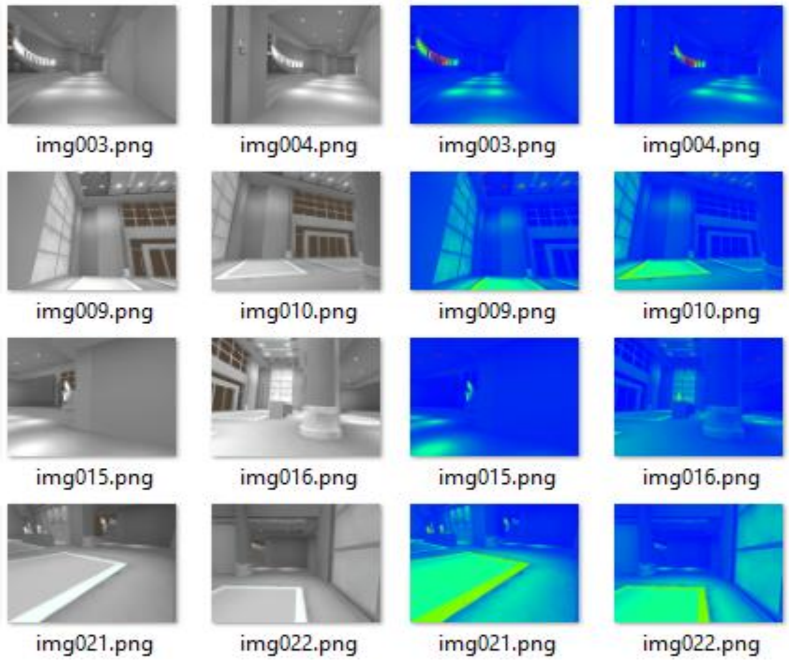

Fig. 2. The training dataset images for light area classification.

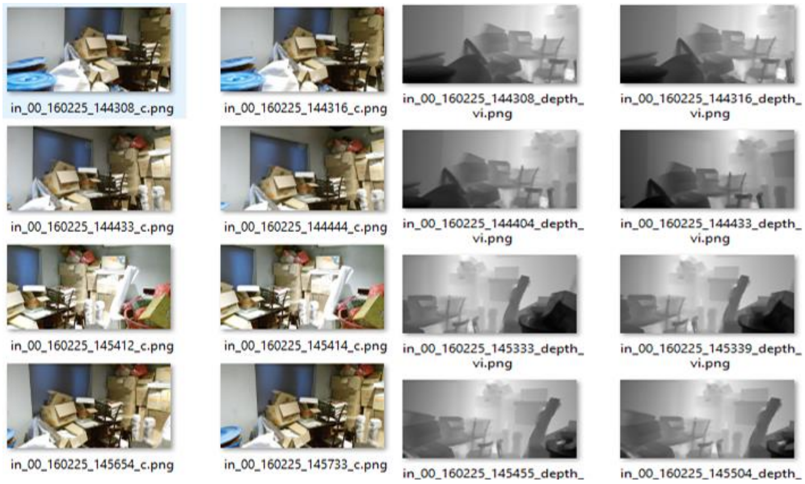
in_00_160225_144308_c.png
in_00_160225_144316_c.png
in_00_160225_144308_depth_vi.png
in_00_160225_144316_depth_vi.png
in_00_160225_144433_c.png
in_00_160225_144444_c.png
in_00_160225_144404_depth_vi.png
in_00_160225_144433_depth_vi.png
in_00_160225_145412_c.png
in_00_160225_145414_c.png
in_00_160225_145333_depth_vi.png
in_00_160225_145339_depth_vi.png
in_00_160225_145654_c.png
in_00_160225_145733_c.png
in_00_160225_145455_depth_
in_00_160225_145504_depth_

Fig. 3. The training dataset images for depth classification.

The history of training and results of work are presented in figures 4 and 5.

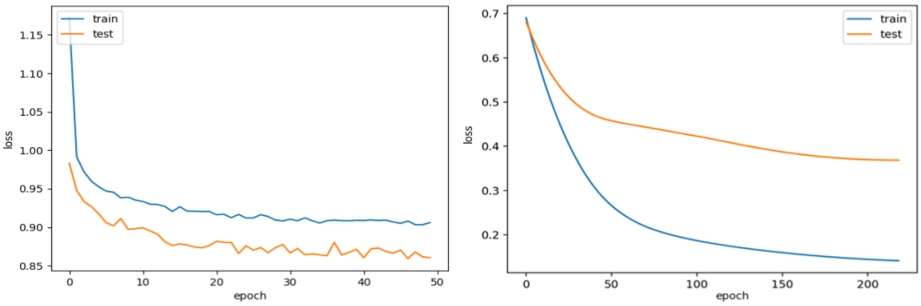

Fig. 4. History of training two VGG16-Nets. In the left - for light detection, in the right – for depth estimation. Epochs are displayed horizontally, and the error of neural network learning is displayed vertically.

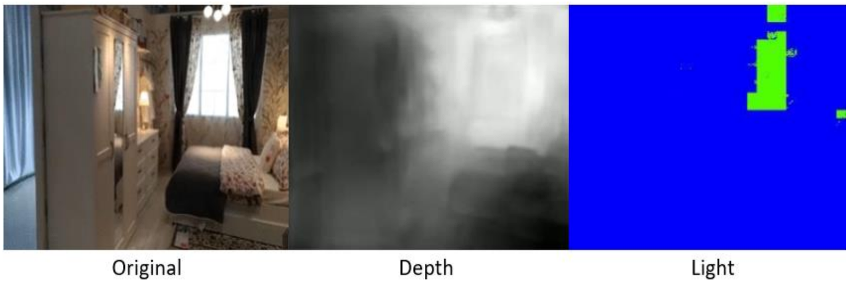

Fig. 5. The results of the trained neural networks with depth and light areas.

4 Conclusion

In this work, neural networks were trained to restore light sources and depth maps of the indoor scene. The architecture of this network is good for classifying data with many classes. The speed allows it to be used in real-time systems. The recognition accuracy of light sources in some scenes turned out to be quite good; in the future, it is planned to improve the coordinate recovery algorithm for greater accuracy, which can easily restore the distance to any point in the image, but working with images significantly faster than with 3D models.

5 Further work

In further work it is planned to improve the results by using other neural network architectures and image processing using computer vision algorithms to strengthen the output results.

References

1. Hold-Geoffroy, Y., Sunkavalli, K., Hadap, S., Gambaretto, E. and Lalonde, J.-F., "Deep outdoor illumination estimation," In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR) (2017).
2. Lalonde, J.-F., Efros, A. A. and Narasimhan, S. G., "Estimating the natural illumination conditions from a single outdoor image," International Journal of Computer Vision, 98(2), 123–145 (2012).
3. Gardner, M.-A., Sunkavalli, K., Yumer, E., Shen, X., Gambaretto, E., Gagne, C. and Lalonde, J.-F., "Learning to predict indoor illumination from a single image," ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia), preprints (2017).
4. Lombardi, S. and Nishino, K., "Reflectance and Illumination Recovery in the Wild," IEEE Transactions on Pattern Analysis and Machine Intelligence 38, 129– 141 (2016).
5. Eigen, D. and Fergus, R., "Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture, " International Conference on Computer Vision (2015).
6. Girshick, R. B., Donahue, J., Darrell, T. and Malik, J., "Rich feature hierarchies for accurate object detection and semantic segmentation," CVPR (2014).
7. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R. and LeCun, Y., "Overfeat: Integrated recognition, localization and detection using convolutional networks," ICLR (2013).
8. Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," CoRR, abs/1409.1556 (2014).
9. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., "Going deeper with convolutions," CoRR, abs/1409.4842 (2014).
10. Heymann, S., Smolic, A., Muller, K., Froehlich, B., "Illumination reconstruction from realtime video for interactive augmented reality," International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS) (2005)
11. Bruno Augusto Dorta, M., Rafael Rego, D., Cristina Nader, Vasconcelos., Esteban, C., "Deep light source estimation for mixed reality," VISIGRAPP 2018 - Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, (303-311) (2018)
12. Salma, J., Philippe, R., Eric, M., "Illumination Estimation Using Cast Shadows for Realistic Augmented Reality Applications," Adjunct Proceedings of the 2017 IEEE International Symposium on Mixed and Augmented Reality, ISMAR-Adjunct (2017)
13. Frahm, Jan-Michael., Koeser, K., Grest, D., Koch, R., "Markerless Augmented Reality with Light Source Estimation for Direct Illumination," European Conference on Visual Media Production, (211-220) (2005)
14. "Lumicept | Integra Inc.," Integra Inc., 2019, <https://integra.jp/en/products/lumicept> (Last accessed 19 Oct 2019).
15. Digital Image Media Laboratory (DIML), “DIML/CVL RGB-D technical_report”, This material is presented to provide a detailed description about the DIML/CVl RGB-D dataset. https://dimlrgbd.github.io/downloads/technical_report.pdf (Last accessed 27 July 2020).
16. Kim, Y., Ham, B., Oh, C., Sohn, K., "Structure selective depth super-resolution for RGB-D cameras," IEEE Trnas. on Image Processing, vol.25, no. 11, pp. 5527-38 (2016).

17. Kim, S., Min, D., Ham, B., Kim, S., Sohn, K., "Deep Stereo Confidence Prediction for Depth Estimation," IEEE International Conference on Image Processing (2017).
18. Kim, Y., Jung, H., Min, D., Sohn, K., "Deep Monocular Depth Estimation via Integration of Global and Local Predictions," IEEE Trnas. on Image Processing, vol.27, no. 8, pp. 4131-43 (2018).
19. Cho, J., Min, D., Kim, Y., Sohn, K., "A Large RGB-D Dataset for Semi-supervised Monocular Depth Estimation", Submitted to the IEEE Transactions on Image Processing (2019).