Recovery of optical parameters of a scene using fully-convolutional neural networks

Maksim Sorokin, Dmitry Zhdanov, Andrey Zhdanov

ITMO University, St. Petersburg, Russia
vergotten@gmail.com, ddzhdanov@mail.ru, andrew.gtx@gmail.com

Abstract. Due to the rapid development of virtual and augmented reality systems the solution of the problem of formation of the natural illumination conditions for virtual world objects in the real environment becomes more relevant. To recover a light sources position and their optical parameters authors propose to use the fully convolutional neural network (FCNN), which allows catching the 'behavior of light' features. The output of the FCNN is a segmented image with luminance levels. As an encoder it was taken the architecture of VGG-16 with layers that pools and convolves an input and wisely classifies it to one of a class which characterizes its luminance. The image dataset was synthesized with use of the physically correct photorealistic rendering software. Dataset consists of HDR images that were rendered and then presented as image in color contours, where each color corresponds to the luminance level. Designed FCNN decision can be used in tasks of definition of illuminated areas of a room, restoring illumination parameters, analyzing its secondary illumination and their classification to one of a luminance level, which nowadays is one of a major task in designing of mixed reality systems to place a synthesized object to the real environment and match the specified optical parameters and lighting of a room.

Keywords: Lighting, FCNN, Segmentation, Deep machine learning.

1 Introduction

This article is devoted to the development of algorithms and methods for solving the global scientific problem in the field of the physically correct and effective restoration of illumination conditions and optical properties of the real-world objects during the synthesis of images of the virtual world. Due to the rapid development of virtual and augmented reality systems the solution of this problem is becoming more relevant. If virtual reality systems were limited only to the visualization of the virtual world objects where the main problem was the

CEUR-WS.org/Vol-2344/short1.pdf

realism of these virtual objects, then in the case of mixed reality systems the new problems arise related to the natural perception of the virtual world objects in the real environments by the device user. One of these problems lies in the area of formation of the natural illumination conditions for objects of the virtual world in the real environment. The actually observed scene may contain different sources of illumination that may be natural, such as the sun or the moon, and artificial, such as lamps or spotlights. The natural illumination conditions can be restored by knowing the geolocation, however the artificial ones must be restored to be clearly defined. In the scope of the current article the main goal is to restore the direct illumination conditions, or in other words the parameters of the light sources, however in some cases, a large role is played by an indirect illumination caused by re-reflection of light from objects of the real world. When a virtual object synthesized in a mixed reality system is mixed to the real world, it should be illuminated naturally, i.e. from the real-world light sources, and it should cast shadow in the direction opposite to the light source. In addition, the virtual object must naturally influence the world around it, i.e. illuminate it with its own or light, reflect in mirrors, shade objects of the real world, etc. 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, as result limiting the time that a person can be in the mixed reality, that in its turn limit the practical use of the mixed reality systems in various areas as for example an education or training.

2 Related works

Currently, convolutional neural networks are actively used for various tasks related directly to an image analysis and processing, be it the classification or recognition of any particular areas. The main advantages of the convolutional neural networks (CNN) include convenient paralleling of computations, resistance to image shift and learning that uses the classical method of backpropagation of error. Among the shortcomings is a large number of adjustable parameters, in particular, setting learning parameters. To solve a problem, one should try various layers and parameters to choose the best solution. These parameters include the convolution kernel dimension, the degree of dimension reduction, the use of subsampling layers, the choice of an activation function, etc. Naturally, the convolutional neural network is well suited for an image segmentation. For example, in [1] a convolutional neural network is used for segmentation of biomedical images. Paper [2] presents a CNN-based technique to estimate a high dynamic range outdoor illumination from a single low

dynamic range image, where the neural network used a large dataset of panoramic images. In [3], convolutional neural networks are used to assess the state of the environment in the open air. By learning on panoramic images and analyzing their parameters such as natural conditions, the geolocation, and brightness of the Sun and sky, the neural network builds a prediction of how the shadow of an object should be placed under specified conditions. Similarly, the work [4] also analyzes objects in the open air. Here the neural network not only determines the position of the sky but also divides the image to halves and analyzing the lower part for shadows build its own predictions. In [5], an analysis of a small area of a room is carried out, which can be used to correctly illuminate the entire visible part of the scene. In paper [6] a similar task was presented that infers object reflectance and scene illumination from an image and recovers its characteristic features, thereby determining whether the scene is indoors or outdoors and reveals properties of materials. In paper [7] EnvyDepth was presented, that creates a detailed collection of virtual point lights that reproduce both the local and the distant lighting effects in the original scene. EnvyDepth produces plausible lighting without visible artifacts, obtaining illumination parameters even in the case of complex scenes, both indoors and outdoors. Another great paper [8] presented their own architecture, addressing three different computer vision tasks using a single basic architecture: depth prediction, surface normal estimation, and semantic labeling, capturing many image details without any super pixels or low-level segmentation. The main difference between the listed papers and the current one is that the neural network designed in the scope of the current paper uses training approach based on the photorealistic rendered images which have known geometry and illumination parameters, and the main goal of the neural network is the recognition of the light source coordinates. Paper [9] presented an approach of understanding a 3D geometry of a scene, by predicting a depth map using a multi-scale deep network. Works [10,11] create semantic segmentations by generating contours and then classifying regions using hand-generated features. Convolutional networks have also been successfully applied in the area of object classification and detection [12, 13, 14, 15]. In most cases, these systems are used to classify either a single object on an image or bounding boxes for several objects of an image. However, CovNets have been applied in different tasks, including pose estimation [16] and stereo depth [17, 18].

3 Implementation

The encoder of FCNN is a part of VGG16-Net architecture. The decoder consists of 3 layers of subsampling as an opposite to convolution, or “transpose convolution” as it is called in another words. The general architecture of the fully convolutional neural network is shown in Figure 1.

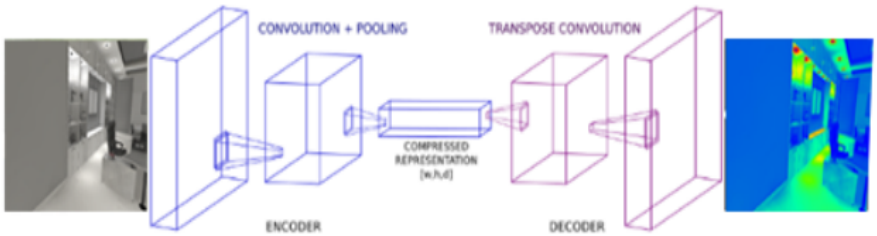

Fig. 1. The architecture of FCNN.

As image dataset was synthesized with use of the physically correct photorealistic rendering software, which has the powerful tools for modeling the light propagation, so there is no doubt of wrong behavior or visualization of primary and secondary illumination, that guarantees proper optical parameters and classification of parameters. The HDR image was rendered and then presented as image in color contours, where each color corresponds to the luminance level. More 'cold' colors mean less intensive illumination and 'hot' colors correspond for the brighter light sources, corresponding to the luminance of a typical room lamp. These images were used to feed CNN to dense layer, where the network learn features to recognize and as output upsamples to a segmentation image. To say more closely about the upsample layer, this is a kind of a function that brings a low-resolution image to a high-resolution one by duplicating each pixel twice, that is also called the nearest neighbor approach. More information is given in paper [20]. On figure 2 you can see the original image with its color contours representation.

As the FCNN is in use, all pixels of input image are classified to one of five classes, which corresponds to the certain color of the color contours representation. As an example, for the color contours representation from the figure 2, the red color specifies that it is the brightest part of the image, and its brightness is about 150 nit, the green color value is near 50 nits, and the blue color means that this part is not illuminated at all. On the figure 3 it is shown the scale

which is measured in the real luminance values and corresponding colors, where 1 nit equals 1 candela per square meter.

Fig. 2. Original image along with its color contours representation.

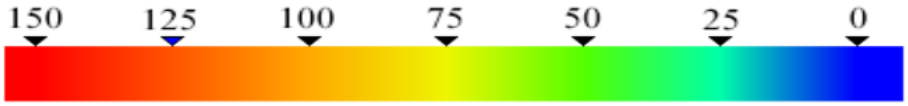

Fig. 3. - Scale of colors and its real luminance values in nits.

But as the color contours representation have many shades and for the classification task it is required that each class corresponds to a specific color, so it was necessary to simplify this representation. On the figure 4 it is shown the image with corresponding black-and-white classes.

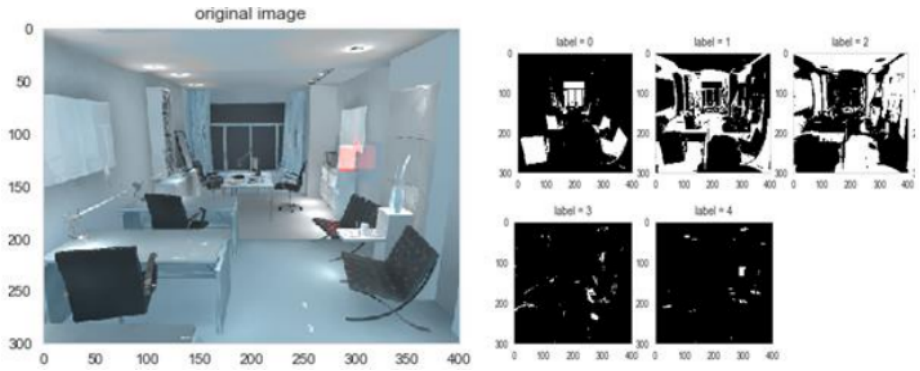

Fig. 4. Image and its classes as black-and-white masks.

After applying the color palette and resizing images, the image dataset looks like the one shown in Figure 5, where left image is a mask and right one is an original image.

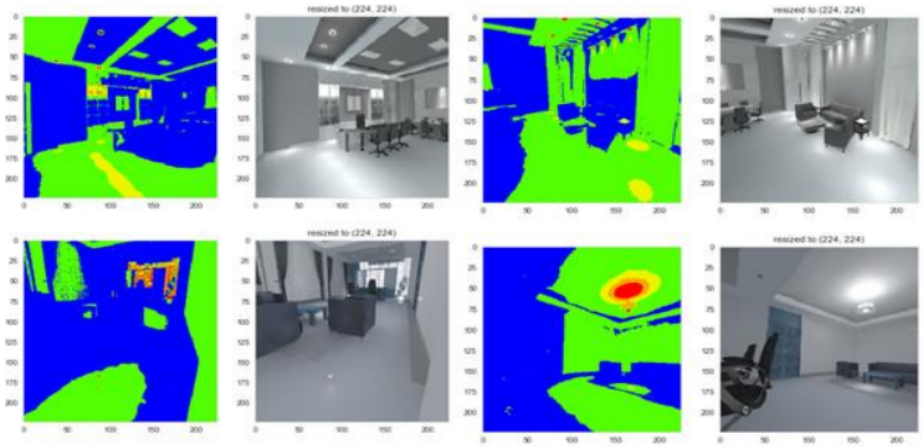

Fig. 5. Image dataset.

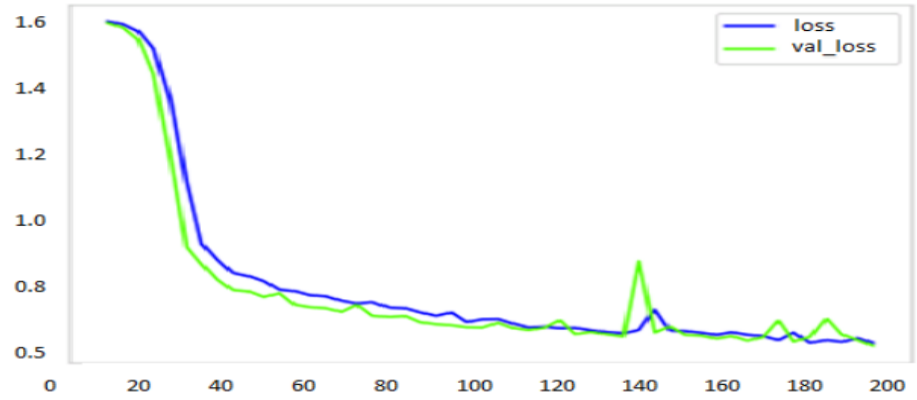

Fig. 6. Training history

Fully-convolutional neural network training was conducted on 221 train images and 29 validation images with learning rate 0.01 and 200 epochs. After training the loss was 0,2. An "intersection over union" method was used as a test, that compares ground truth area of an input image and output image, comparing its pixels and giving the accuracy as result. The mean intersection is 0.7, almost

rightly classifying the first class with a value of 90 percents of accordance and the last class with a probability of 30 percents. The figure 6 shows the history of training and the figure 7 shows the obtained results, where original image is to the left, the predicted result is in the middle and the ground truth is to the right. Lately, this method will be expanded not only to restore the illuminated areas, but also to determine the light sources and their position and direction.

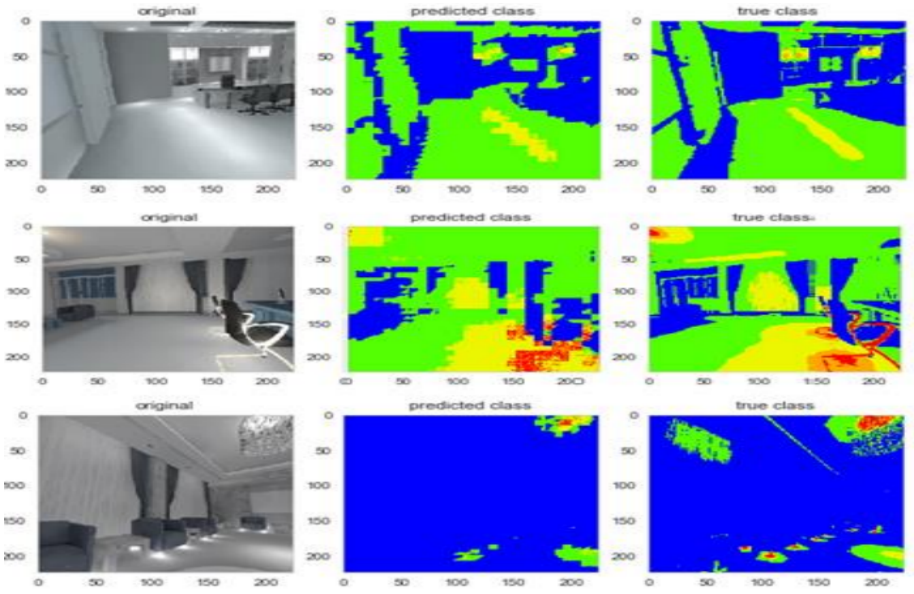

Fig. 7. The obtained results of FCNN.

4 Conclusion

In the scope of the current research the significant improvement in the field of the illumination conditions analysis was achieved. The dataset of illumination fully-convolutional neural network contains 291 images with their color contours representation. The neural network can classify the input image to certain classes of illumination and segment the image as output, but there still exists a problem of classifying the diffuse and mirror reflections. The continuation of the research will be aimed on the improvements in classifying not only the luminance values but also the illuminance values and bounding boxes of light sources, that can highlight them more efficiently. Now FCNN decision can be used in tasks of the definition of illuminated areas of the environment, restoring

luminance parameters, taking features of shadows, analyzing secondary illumination and classifying them to one of luminance levels. Nowadays it is one of the major tasks in mixed reality systems design, required to place the synthesized virtual objects to the real environment and match the existing light-optical characteristics of the real environment. Speaking about the determination of light, the CNN encoder can determine the type of illumination, if it is the ceiling ones or wall light sources. Considering the fact that all dataset images are synthesized, and the geometry of each scene is well known, we can confidently correlate the known coordinates with the predicted ones to create the more complete picture.

Acknowledgements

This work was funded by Russian Science Foundation (project No. 18-79-10190).

References

1. Ronneberger O., Fischer P., Brox T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N., Hornegger J., Wells W., Frangi A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, vol 9351. Springer, Cham (2015).
2. 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).
3. 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).
4. Gardner M.-A., Sunkavalli K., Yumer E., Shen X., Gambaretto E., Gagné C., and Lalonde J.-F.: Learning to predict indoor illumination from a single image. ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia), preprints (2017).
5. Hinton G.E., Nair V.: Rectified linear units improve restricted boltzmann machines // In Proceedings of the 27th International Conference on Machine Learning, 807–814 (2010).
6. Lombardi S. and Nishino K.: Reflectance and Illumination Recovery in the Wild. IEEE Transactions on Pattern Analysis and Machine Intelligence 38, 129– 141 (2016).
7. Banterle F., Callieri M., Dellepiane M., Corsini M., Pellacini F., and Scopigno R.: EnvyDepth: An interface for recovering local natural illumination from environment maps. Computer Graphics Forum 32, 7, 411–420 (2013).

8. 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).
9. Eigen D., Puhrsch C., and Fergus R.: Depth map prediction from a single image using a multi-scale deep network. NIPS (2014).
10. Gupta S., Arbelaez P., and Malik J.: Perceptual organization and recognition of indoor scenes from rgb-d images. In CVPR (2013).
11. Gupta S., Girshick R., Arbelaez P., and Malik J.: Learning rich features from rgb-d images for object detection and segmentation. In ECCV (2014).
12. Girshick R. B., Donahue J., Darrell T., and Malik J.: Rich feature hierarchies for accurate object detection and semantic segmentation. CVPR (2014).
13. Sermanet P., Eigen D., Zhang X., Mathieu M., Fergus R., and LeCun Y.: Overfeat: Integrated recognition, localization and detection using convolutional networks. ICLR (2013).
14. Simonyan K. and Zisserman A.: Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556 (2014)
15. 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).
16. Tompson J., Jain A., LeCun Y., and Bregler C.: Joint training of a convolutional network and a graphical model for human pose estimation. NIPS (2014).
17. Zbontar J. and LeCun Y.: Computing the stereo matching cost with a convolutional neural network. CoRR, abs/1409.4326 (2014).
18. Memisevic R. and Conrad C.: Stereopsis via deep learning. In NIPS Workshop on Deep Learning (2011).
19. Lumicept Homepage, https://integra.jp/en/products/lumicept, last accessed 2019/03/14.
20. Long J., Shelhamer E., Darrell T.: Fully Convolutional Networks for Semantic Segmentation. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431-3440 (2015).