Towards fast lighting condition inference for augmented reality

Leon Modic and Luka Čehovin Zajc

University of Ljubljana, Faculty of computer and information science
E-pošta: lm4903@student.uni-lj.si, luka.cehovin@fri.uni-lj.si

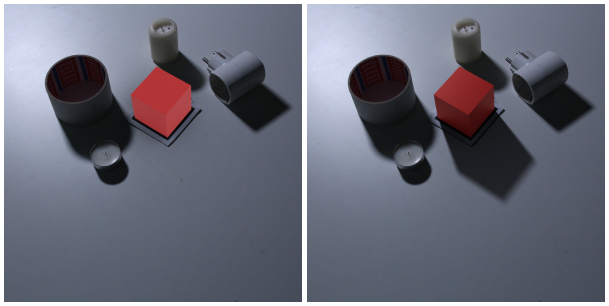
Abstract

Augmented reality (AR) merges real world with digital content, a process which requires information about the observed scene. In this paper we address a part of this task by exploring fast inference of a single light setup in common AR scenarios using deep learning. We propose a new synthetic dataset for training deep models for the task of light condition inference as well as a dataset of photos used for testing models in real environment. Using our datasets we have compared existing approaches that vary in light position information representation. Additionally, we propose an alternative representation that could extend to encoding multiple lights. We discuss the differences between evaluated models and provide some ideas for further research in this field.

1 Introduction

Augmented reality aims to blend real world with digital content which requires information about the scene, primarily its structure and camera position. One of additional visual cues that makes the superimposed digital content appear more realistic are matching lighting conditions. While lighting is not necessary for all AR applications (e.g. information overlays), other applications like gaming may benefit from it. In these scenarios lighting may make or break the illusion as illustrated in Figure 1.

Real-world scene lighting is a complex phenomenon, but the parameters that are inferred for AR are firstly limited by the real-time graphic engines which operate with a number of primitive light sources with different intensity. Secondly, these parameters have to be further simplified due to real-time and low power consumption constraints. To this end, most research work in this area focuses on determining position of a single light source relative to the camera using regression of angle parameters. In this paper we continue the investigation by proposing a new synthetic dataset for training deep models and a real dataset for evaluation. We also show that superior performance can be achieved by treating the problem as a classification problem where individual classes denote discretized angles. We believe that this formalization can be easily extended to multiple light sources.



(a) (b)

Figure 1: Comparison of lighting setups. Image (a) shows a generic render of a cube where no information about the light is known. Image (b) shows a render of the same cube with light conditions matching those in the scene. Both examples are rendered using a fast OpenGL renderer.

2 Related Work

There have been several methods for illumination estimation proposed in the past years, each with different assumptions and goals. Most early approaches utilize light probes and/or objects with known geometry, e.g. in [1] an object with known geometry and Lambertian reflectance material assumption is used to predict position of multiple light sources. Light probes are used in AR context in [2], in [3] a known, globally convex, object is also used as a light probe to determine position and intensity of a light. Combining both ideas [4] proposes a deep learning model trained on synthetic versions of a known model that is used as a light probe in to determine lighting conditions in real images. In [5] authors propose an interactive method for light estimation without known geometry, but require user to trace the silhouette of an object. In [6], authors use head-mounted camera to classify lighting conditions into a predefined number of illumination classes based on illumination of hands.

Direct parameter inference using deep models has been recently explored by [7]. Synthetic RGB-D images are used to train prediction of direction of a single light in an AR scene. Their work has been extended by [8], they work with RGB images and use stereoscopic projection to encode light direction. Our work follows these recent developments. We use synthetic RGB data for training the model and a real dataset for testing. In contrast to [8] our datasets are more diverse and contain more than

Figure 2: Examples from synthetic training (above) and real-world testing (below) datasets.

one object in various spatial configurations. We are also proposing a new way to encode light position that has potential to generalize to multiple light sources.

3 Datasets

As mentioned, recent works [7, 8] utilize synthetic datasets for training their models, but they have so far not released their datasets to the research community. The datasets are also quite limited in diversity of scenes. We have therefore built our own dataset to address these issues. The dataset is generated in Blender using Cycles ray-tracing engine. Scenes, materials, light and camera position are generated using pseudo random generator which means that rendering of specific sample can be reproduced by controlling the seed value. At the same time the size of the dataset can be easily extended by generating new samples. For the experiment, presented in this paper, we have generated a dataset consisting of more than 60.000 samples using a single light source in combination with ambient light to mimic the diversity of real-word conditions. Several images from the dataset are shown in Figure 2.

The second dataset was acquired for testing. The acquisition was done using a Canon EOS700D camera in a controlled environment with a single dominant light source. The position of a camera and position of the light source was monitored determined using fiducial markers. By also changing objects in the scene, we have acquired 100 images, some of those shown in Figure 2. We have also created an online annotation tool that allows us to

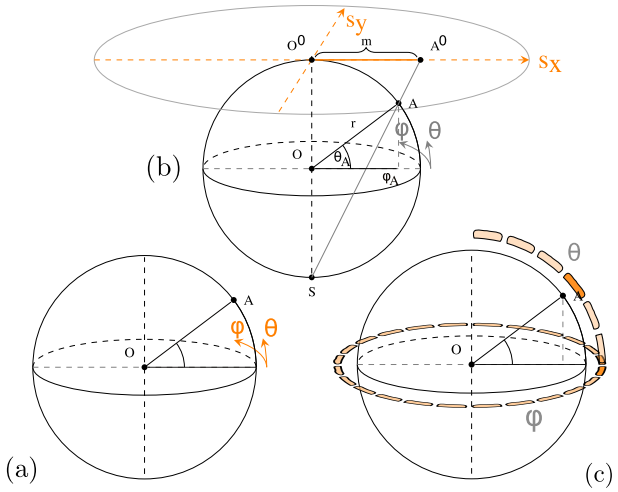


Figure 3: Light angle encoding approaches. (a) Direct encoding using radians - RAD, (b) Stereographic projection - STR, (c) Angle discretization - DIS.

manually to determine position of light using slider controls and a visualization of an shaded object as a feedback. We intend to use this tool to annotate in-the-wild samples. So far we have used it to re-annotate half of the testing dataset in order to evaluate human precision for the same task that is expected from evaluated models.

4 Methodology

Based on our previous preliminary experiments we have selected EfficientNet [9] architectures for all our experiments with the proposed dataset. The model receives a RGB image of 128×128 pixels as an input. For the output we have first implemented two angle regression approaches, regression of radian angles of light relative to the camera [7], denoted as RAD, and stereographic projection of the angles [8], denoted as STR. We have relied on the description in corresponding papers and only changed the backbone part.

Both reference approaches are illustrated in Figure 3 and make certain assumptions regarding the scene that make sense in AR scenarios. The φ angle denotes the azimuth angle and spans from 0 to 360 degrees, while θ denotes elevation angle and only spans from 0 to 90 degrees. The advantage of stereographic projection of angles is supposed to disentangle interaction of both angles when the θ approaches 90 degrees and estimation of φ becomes difficult.

The reference approaches [7, 8] have a problem that they only model a single light source. Every extension of the model to multiple lights or addition of ambient light requires modification of the architecture. There is also a problem of selecting appropriate loss function that has to take these interdependent factors into account. We are therefore exploring an option of formalizing lighting conditions as a discretized probability function. For a single light source the problem can be reduced to a separable multi-class problem where the interval of possible angle values is linearly discretized to a certain number of options. The desired output is then encoded using one-hot

	Model		Random (1)		ImageNet (2)	
	R	A	φ	θ	φ	θ
A	RAD	B3	$30,5 \pm 32,8(21,2)$	$20,5 \pm 13.3(19,1)$	$24,4 \pm 22,4(16,9)$	$23,6 \pm 11,4(24,4)$
B	STR	B3	$62,8 \pm 48,9(49,2)$	$20,1 \pm 12,6(20,9)$	$51,5 \pm 46,5(28,3)$	$18,9 \pm 12,1(17,8)$
C	DIS	B3 (64, 64)	$37,7 \pm 41,4(22,2)$	$21,5 \pm 14,5(19,3)$	$23,8 \pm 29,3(14,6)$	$25,6 \pm 11,7(24,4)$
D	DIS	B3 (32, 16)	$19,0 \pm 21,8(12,0)$	$19,6 \pm 12,2(19,0)$	$20,3 \pm 17,3(15,5)$	$22,6 \pm 10,3(22,4)$
E	DIS	B0 (32, 16)	$29,9 \pm 32,9(19,3)$	$21,1 \pm 12,6(19,9)$	$21,5 \pm 26,4(14,9)$	$22,1 \pm 10,8(22,2)$

Table 1: Overview of the experiment results. Each model is represented with a letter from A to E and a number 1 or 2. Model hyperparameters are defined with angle representation (R) and architecture (A). For discretized angles approach we have evaluated several configurations, two different discretization resolutions and two different models (EfficientNet B3 and B0). Number presented are angle errors on testing dataset (smaller is better) converted to degrees. The first number is average error, followed by variance, the third number is median error.

encoding. Note that this approach introduces some level quantization noise. This noise could be mitigated using interpolation, but at the moment we have only evaluated this simple approach.

5 Results

We have evaluated all described approaches in two different contexts, we have used EfficientNet B3 as the basis, once with randomly initialized parameters, in second case model parameters were initialized using a model pre-trained on ImageNet dataset. All models were trained using Adam [10] optimization algorithm. Learning rate was set to 0.0002 for randomly initialized models variants and to 0.0001 for the pre-trained variants. Batch size was set to 32 samples. According to [7, 8] we used MSE loss function for RAD and STR approaches. For DIS, we have used standard cross entropy loss. We have run training for all the models for 100 epochs.

The results of our experiment are presented in Table 1, all errors are given in degrees. The errors for all models look quite high, but to put the numbers into perspective we have measured average error of human annotations for a part of the testing dataset which are 24.9 for φ and 7.3 for θ. It is also clear that the inter-sample variation of error is quite high. We have therefore also computed median angle error which is almost always lower than the average. This shows that the distribution is skewed - we have many samples with low error and a few samples with very high error as shown in Figure 4 where we look at the distribution of errors for model D1. The errors according to φ are indeed not distributed normally. The θ errors are more evenly distributed and are apparently harder to infer.

Reference models (A and B) are performing worse than reported in corresponding papers [7, 8], also due to a more difficult setup with more objects in the scene as well as textures. The best models according to the testing set results are the discretized position models, primarily D1 and D2, despite quantization errors that occur in this representation. Increasing discretization resolution (models C1 and C2) does not help, but increases error, especially with respect to φ dimension. We attribute this to the simple quantization technique with one-hot encoding, we believe that distributing *votes* among neighbor cells may reduce this problem. Selected examples of predictions for D1 model are shown in Figure 5, first row shows good examples, second row bad ones. We can see that the model is less sure regarding the θ angle with more uncertainty in the distribution. We can also see that model tends to be less reliable when real objects are not distributed over entire image. Another thing that we have observed is that the model performs worse on the half of images that contain fiducial markers. All these observations give us new guidelines and ideas for the design of the training dataset.

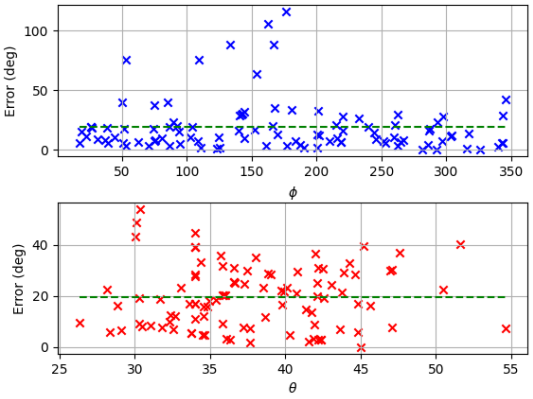


Figure 4: Distribution of errors for testing dataset for angle φ and θ according to model D1. The plots show errors for all testing samples with respect to their actual angle. Dotted green lines represent average error for each distribution.

Finally, we have tested out trained model on several images that we have acquired in-the-wild with more diverse and different light sources. Figure 6 contains several examples where the model behaves well despite weak shading and reflections, leading us to believe that a more general and robust model can be obtained with more work.

6 Conclusion

In this paper we have presented our preliminary results on fast inference of a single light setup in common AR scenarios using deep learning. We have proposed a new synthetic dataset for training deep models for the task of light condition inference as well as a dataset of real images used for testing models in real environment. Both datasets are meant to become publicly available, thus benefiting the progress of the field. Using these datasets we have evaluated several approaches and proposed a novel representation that could be easily extended to encode

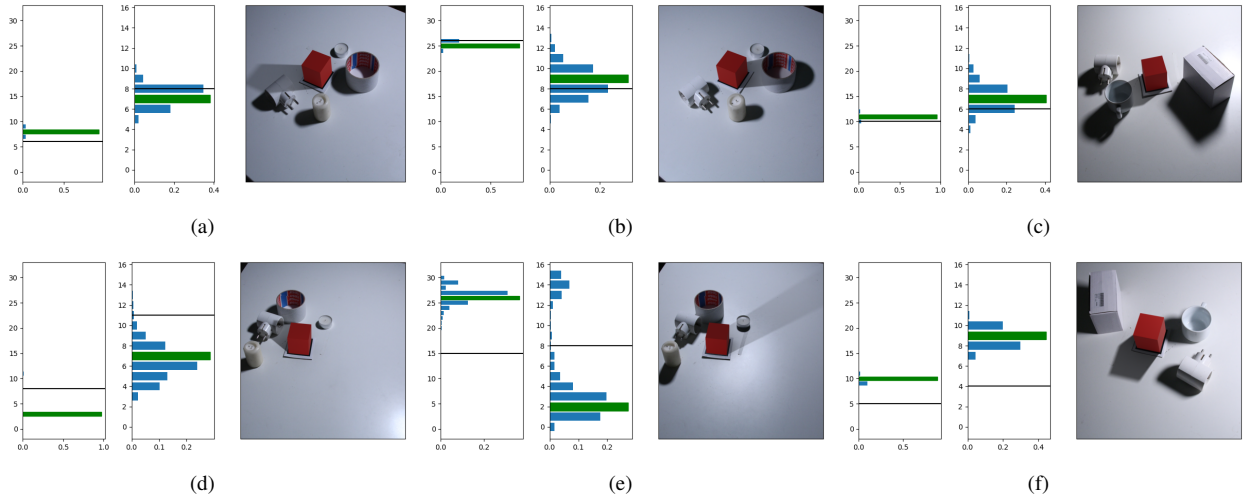


Figure 5: Qualitative example overview for model D1. Prediction distribution for φ is presented on the left chart and for θ on the right chart. Examples (a) to (c) show results with good prediction of both angles, examples (d) to (f) show cases with severe problems. Reference object rendered using an OpenGL based engine, shadow intensity a default value and it not predicted by the model.

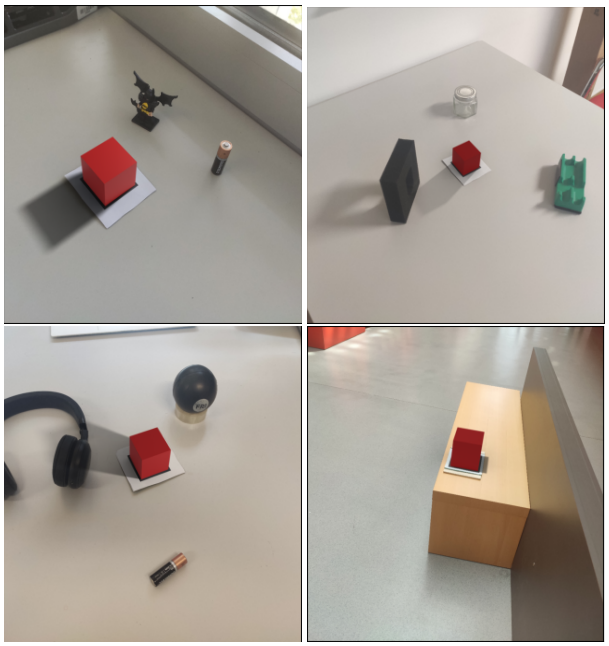


Figure 6: Preliminary result for D1 model on samples recorded outside the controlled environment. Only position is determined by the model, other light properties are set to default values.

multiple lights. Our future work will involve improving and extending the datasets as well as exploring ways to determine multiple light sources in a scene and working with multiple viewpoints to increase robustness.

Acknowledgements: This research was partially funded by ARRS research program P2-0214.

References

[1] Yang Wang and Dimitris Samaras. Estimation of multiple directional light sources for synthesis of augmented reality images. *Graphical Models*, 65(4):185–205, July 2003.

[2] K. Agusanto, Li Li, Zhu Chuangui, and Ng Wan Sing. Photorealistic rendering for augmented reality using environment illumination. In *The Second IEEE and ACM International Symposium on Mixed and Augmented Reality, 2003. Proceedings.* IEEE Comput. Soc, 2003.

[3] Rang M. H. Nguyen and Minh Ngoc Le. Light source estimation from a single image. In *2012 12th International Conference on Control Automation Robotics & Vision (ICARCV)*. IEEE, December 2012.

[4] David Mandl, Kwang Moo Yi, Peter Mohr, Peter M. Roth, Pascal Fua, Vincent Lepetit, Dieter Schmalstieg, and Denis Kalkofen. Learning lightprobes for mixed reality illumination. In *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, October 2017.

[5] Jorge Lopez-Moreno, Sunil Hadap, Erik Reinhard, and Diego Gutierrez. Light source detection in photographs. In *CEIG*, pages 161–167, 2009.

[6] Bruno Augusto Dorta Marques, Rafael Rêgo Drumond, Cristina Nader Vasconcelos, and Esteban Walter Gonzalez Clua. Deep light source estimation for mixed reality. In *VISIGRAPP*, 2018.

[7] Peter Kán and Hannes Kafumann. DeepLight: light source estimation for augmented reality using deep learning. *The Visual Computer*, 35(6-8):873–883, May 2019.

[8] Markus Miller, Alfred Nischwitz, and Rüdiger Westermann. Deep light direction reconstruction from single RGB images. In *WSCG 2021 Proceedings*, 2021.

[9] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[10] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.