Single Image Multi-Spectral Photometric Stereo Using a Split U-Shaped CNN

Doris Antensteiner, Svorad Štolc and Daniel Soukup
AIT Austrian Institute of Technology GmbH
Vienna, Austria
firstname.lastname@ait.ac.at

Abstract

We present a system to extract surface orientation and albedos from a single shot image using three differently colored illumination sources. Photometric stereo allows one to extract local surface information such as normals or gradients. Traditionally, the local orientations and albedos are computed using serveral acquisitions of the same viewing angle and under varying illumination directions. In applications with moving objects, where the acquisition- as well as processing speed are essential, such setups are poorly suited. We propose a single shot decomposition using three differently colored light sources under defined illumination directions. To allow for a fast and regularized inference, we built a split U-shaped convolutional neural network, which takes a single shot input and estimates both the surface orientation and albedo simultaneously.

1. Introduction

Estimating physical properties such as the local surface orientation and albedo of a scene from a single image is a challenging problem. We propose a method to infer such underlying parameters for real world data with defined illumination directions and colored (RGB) light sources using a split U-shaped convolutional neural network (split U-Net).

Photometric stereo (PS) methods as presented in [1] recover surface orientation and allow the reconstruction of high frequency surface details. Such data can be used to partly reconstruct depth information [2], up to the bas-relief transform [3].

A method to find the most likely surface orientation and reflectance under unknown illuminations from a single image was previously demonstrated in [4]. Later, a refined method for three calibrated (RGB) monochromatic directional light sources was shown in [5]. Surface orientations and reflections were recovered based on assumptions such as a piece-wise constant texture and a finite set of distinct albedo values. Learning frameworks for photometric stereo were presented in [6, 7, 8], where images captured under different illumination directions were utilized to predict a normal map. U-shaped networks were previously used in various ways, including the analysis of image stacks in the form of light field data [9].

In our work, we use shape and reflection priors under known light configurations to extract the albedo, surface orientations and subsequentially depth data from a single shot image. We designed and used a split U-Net for the task of inferring such data while regularizing the output. The network is trained on real datasets, which are not strictly Lambertian. Contrary to previous contributions, we designed and trained a novel split U-Net for which we used ground truth (GT) data, calculated by illuminating the object from multiple (32) directions.

2. Photometric Stereo

On a discretized surface with the size of $M \times N$ pixels with Lambertian reflectance, the surface normals $\boldsymbol{N}_{i,j} \in \mathbb{R}^3$, for all pixel locations $(i,j) \in M \times N$, and the albedo $\rho_{i,j} \in \mathbb{R}$ are reconstructed under defined illumination sources $\boldsymbol{L} \in \mathbb{R}^{n\times 3}$. The observed n intensities are defined as $\boldsymbol{E}_{i,j} \in \mathbb{R}^n$. The following tensors hold vectors in each pixel location and are denoted with bold characters:

$$\boldsymbol{M}_{i,j} = \rho_{i,j}\boldsymbol{N}_{i,j}, \quad (1)$$

$$\boldsymbol{M}_{i,j} = (M_{i,j,x}, M_{i,j,y}, M_{i,j,z}), \quad (2)$$

$$\boldsymbol{N}_{i,j} = (N_{i,j,x}, N_{i,j,y}, N_{i,j,z}), \quad (3)$$

$$\boldsymbol{E}_{i,j} = (E_{i,j,1}, ..., E_{i,j,n}). \quad (4)$$

Normals and albedos are recovered from observed intensities by using the following least squares (LS) formulation:

$$\min_{\boldsymbol{M}_{i,j}} \frac{1}{2}||\boldsymbol{L}\cdot\boldsymbol{M}_{i,j} - \boldsymbol{E}_{i,j}||^2. \quad (5)$$

The length of the vector $\boldsymbol{M}_{i,j}$ is defined by the albedo $\rho_{i,j}$, as per definition normals are unit vectors:

$$\sqrt{N_{i,j,x}^2 + N_{i,j,y}^2 + N_{i,j,z}^2} = 1, \quad (6)$$

$$\rho_{i,j} = \sqrt{M_{i,j,x}^2 + M_{i,j,y}^2 + M_{i,j,z}^2}. \quad (7)$$

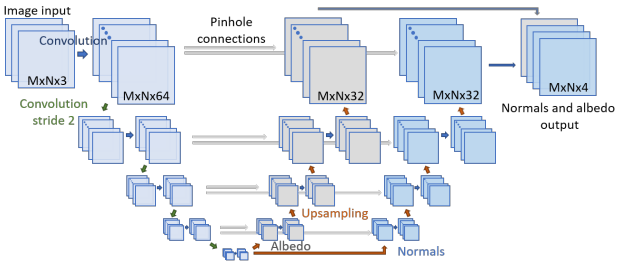

Figure 1: Structure of our proposed split U-Net with pinhole connections. The encoder is represented on the left and the decoder on the right.

Decomposing and analyzing single images from acquisitions with three colored (RGB) light sources to extract variable albedos and surface normals is a challenging problem. The reflectance analysis over our colored channels allows ambiguous interpretations, as specific radiance values can stem from the shape / scaling, light source transformations or texture changes of the object.

3. Network Architecture

We present a split U-Net, which takes a single image as input and jointly estimates the albedo as well as the surface normal, as shown in Fig. 1. This network architecture allows a fast inference due to the data compression, while enabling a detailed reconstruction on fine scales. The image was taken under three colored illuminations in a single shot, matching the sensitivity peeks in the camera sensor. The split U-Net consists of an encoder path and a decoder path. We split the network at bottleneck (between encoder and decoder) and separate the decoder, which allows the network to achieve a more specialized inference for both independend sub-problems. Pinhole connections are established between the encoder and the decoder at each level.

3.1. Loss Function

A loss function is specified in order to minimize the difference between the network output ($\boldsymbol{N}$, ρ) and our GT data ($\hat{\boldsymbol{N}}$, $\hat{\rho}$), respectively. It is defined as the sum of three loss functions, namely the albedo loss $\mathcal{L}_A$, the cosine distance $\mathcal{L}_N$ between the normals and $\mathcal{L}_U$, which enforces unit length normal vectors.

$$\mathcal{L} = \frac{1}{MN}\sum_{i}^{M}\sum_{j}^{N}(\mathcal{L}_{A_{i,j}} + \mathcal{L}_{N_{i,j}} + \mathcal{L}_{U_{i,j}}) \quad (8)$$

$$\mathcal{L}_{A_{i,j}} = (\rho_{i,j} - \hat{\rho}_{i,j})^2 \quad (9)$$

$$\mathcal{L}_{N_{i,j}} = \left(1 - \frac{\langle \boldsymbol{N}_{i,j}, \hat{\boldsymbol{N}}_{i,j}\rangle}{||\boldsymbol{N}_{i,j}||||\hat{\boldsymbol{N}}_{i,j}||}\right)^2 \quad (10)$$

$$\mathcal{L}_{U_{i,j}} = (||\boldsymbol{N}_{i,j}|| - 1)^2 \quad (11)$$

[MSE]		RGB U-Net (ours) ρ	RGB U-Net (ours) N	3L-PS (ref. method) ρ	3L-PS (ref. method) N
32L-PS GT	train	**0.00096**	**0.00146**	0.08933	0.00312
	validation	**0.00118**	**0.00102**	0.09093	0.00390
	test	**0.00148**	**0.00353**	0.06832	0.00371

Table 1: Quantitative evaluation of the distance to the photometric stereo GT (32L-PS) to our RGB U-Net results (bold) in comparison to the reference method using 3 light sources (3L-PS) and a LS algorithm.

The loss function $\mathcal{L}$ was optimized with the Adam optimizer [10] with an initial learning rate of $l = 10^{-4}$ with a decay rate of 10^{-1} each 1000 iterations.

4. Dataset

A light dome with 32 illumination sources was used to acquire our dataset, the system was previously described in [11]. The GT ($\hat{\boldsymbol{N}}$, $\hat{\rho}$) was calculated using all available illumination sources (32L-PS). We created our network input images using three illumination sources with an angular distance of 120° to each other, well aligned to the data used to calculate the GT. We composed one RGB photometric image using three light sources by constructing the R, G and B channel from the first, second and third acquired image under white light respectively. In the training set 78 samples are used, 20 samples in the validation set and 28 samples in the test set.

5. Evaluation

To evaluate our split U-Net, which estimates the albedo and surface normals from a one-shot image with three colored illumination sources (RGB U-Net), we compare it to our GT data computed as described in Sec. 2. We generated this GT data using 32 illumination sources (GT 32L-PS). Additionally, we compare the prediction by our neural network (RGB U-Net) with a least squares (LS) estimation using the same 3 illumination sources (3L-PS). Further than learning the albedo and surface normals, the CNN acts as a regularizer. This becomes obvious in areas containing noise and outliers, which are especially present in dark regions, e.g. on the background plane of our samples (see Fig. 2).

We performed quantitative evaluations of the distance between our GT (32L-PS) to the train, validation and test data respectively. As shown in Tab. 1, the mean squared error (MSE) of our trained network (RGB U-Net) is closer to the ground truth data (32L-PS) than the LS result using the same input (3L-PS).

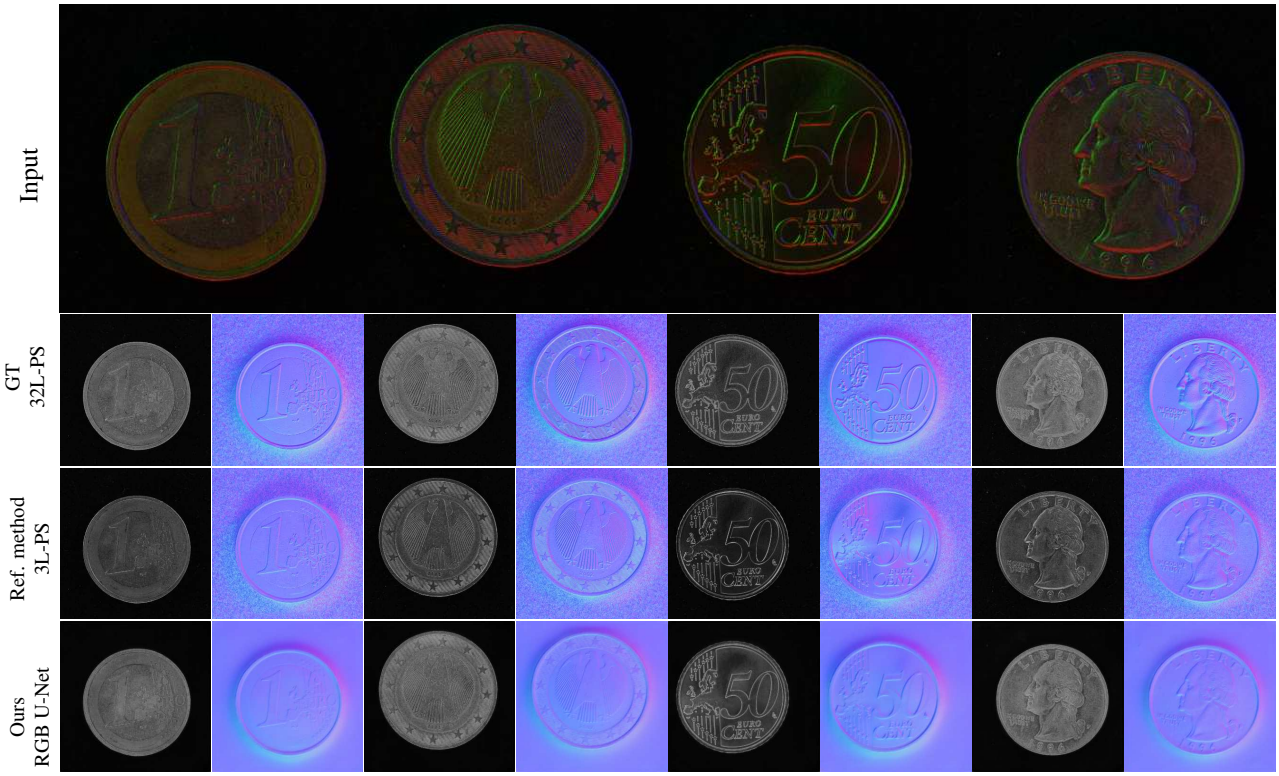

Figure 2: Examples of input images are shown in the first row. The first two samples were seen during the training, the last two were in the test set. The corresponding ground truth data (calculated using 32 illuminations) is shown in the second row. The third row demonstrates the LS reference approach and the result estimated by our network is shown in fourth row.

6. Conclusions

We presented a system to estimate suface normals and albedos from single shot images using colored illumination sources (Fig. 2, row 1). The labels for our training data were obtained by a standard strobing approach using additional light sources (32) to estimate a precise GT (row 2). The quadratic energy function was optimized using a LS solver. Seeing only 3 light sources at the input, our system could estimate a more reliable surface structure (row 4) than the comparison LS reference algorithm (row 3). Using the proposed split U-Net allows a fast inference, which is crucial for industrial applications.

References

[1] R. J. Woodham, "Photometric method for determining surface orientation from multiple images," vol. 19. International Society for Optics and Photonics, 1980.

[2] R. T. Frankot and R. Chellappa, "Method for enforcing integrability in shape from shading algorithms." vol. 10, no. 4, 1988, pp. 439–451.

[3] P. N. Belhumeur, D. J. Kriegman, and A. L. Yuille, "The bas-relief ambiguity," ser. CVPR. IEEE Conference on Computer Vision and Pattern Recognition, 1997.

[4] J. T. Barron and J. Malik, "Shape, illumination, and reflectance from shading," vol. 37, no. 8. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, pp. 1670–1687.

[5] A. Chakrabarti and K. Sunkavalli, "Single-image RGB photometric stereo with spatially-varying albedo," vol. abs/1609.04079, 2016.

[6] G. Chen, K. Han, and K. K. Wong, "PS-FCN: A flexible learning framework for photometric stereo," vol. abs/1807.08696, 2018.

[7] H. Santo, M. Samejima, Y. Sugano, B. Shi, and Y. Matsushita, "Deep photometric stereo network." IEEE International Conference on Computer Vision Workshops (ICCVW), 2017, pp. 501–509.

[8] T. Taniai and T. Maehara, "Neural photometric stereo reconstruction for general reflectance surfaces," vol. abs/1802.10328, 2018.

[9] S. Heber, W. Yu, and T. Pock, "U-shaped networks for shape from light field," E. R. H. Richard C. Wilson and W. A. P. Smith, Eds., 2016, pp. 37.1–37.12.

[10] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," vol. abs/1412.6980. CoRR, 2014.

[11] S. Štolc, D. Soukup, and R. Huber-Mörk, "Invariant characterization of dovid security features using a photometric descriptor." IEEE International Conference on Image Processing (ICIP), 2015, pp. 3422–3426.