An Introduction to Benchmark Dataset

Hongzhi Liu

April 15, 2018

1 Benchmark Dataset for Evaluating Single-Image

In the last reading article, I presented a brief introduction about single-image reflection removal algorithms which aim to remove undesired reflections from a photo. And I will continue to learn about an important part of this theory in PhD Wan's thesis called benchmark dataset and the method of evaluating single-image.

An overview of the scenes in SIR2 dataset is in Figure 1 and the dataset has four major characteristics. With ground truth provided, the team treated a triplet of images as one set, which contains the mixture image, and the ground truth of background and reflection. They created three sub-datasets: The first one contains 20 controlled indoor scenes composed by solid objects; the second one uses postcards to compose another set of 20 different controlled scenes; and the third one contains 100 different wild scenes. For each triplet in the controlled scene dataset, They took images with 7 different DoFs (by changing the aperture size and exposure time) plus 3 different thicknesses of glass. In total, the dataset contains $(20 + 20) \times (7 + 3) \times 3 + 100 \times 3 = 1500$ images.

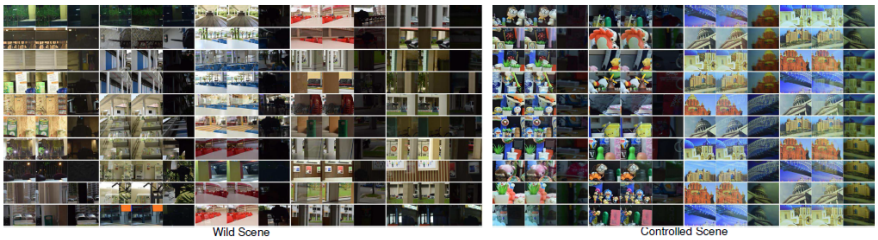

Figure 1: An overview of the SIR2 dataset: Triplet of images for 50 (selected from 100, see supplementary material for complete examples) wild scenes (left) and 40 controlled scenes (right).

Moreover, the misalignment among multiple images might prevent these methods from being applied to scenes in the wild and mobile devices. The benchmark dataset and evaluating single-image methods have consistently simple request for input data capture and great potential for wide applicability.

2 Reflection Removal Algorithms Evaluation

The team used the SIR2 dataset to evaluate representative single-image reflection removal algorithms for both quantitative accuracy and visual quality. They chose four methods, because those are recent methods belonging to different types with state-of-the-art performance.

For each four evaluated method, default parameters had been used to be suggested in papers or used in their original codes. AY07 [17] requires the user labels of background and reflection edges, so it needs to be followed the guidance to do the annotation manually. SK15 [29] requires a pre-defined threshold (set as 70 in their code) to choose some local maxima values. However, such a default threshold shows degenerated results on dataset, and they manually adjust this threshold for different images to make sure that a similar number of local maxima values to their original demo are generated. To make the image size compatible to all evaluated algorithms, researchers resized all images to 400×540.

Through reading this thesis, I can learn the way of thinking by following their research steps. And quantitative evaluation in the next section is performed by checking the difference between the ground truth and the estimated from four methods which are mentioned above. I will try to understand tables and formulas to know the evaluation process.