IST5-3L
Late-News Paper

Estimating Scene Luminance Levels from Images Using Deep Learning

Ya-Hsuan Lin[1], Pei-Li Sun[1*], I-Chin, Wu[2], Min Di[2]

m11225014@mail.ntust.edu.tw, *plsun@mail.ntust.edu.tw
[1] Graduate Institute of Color and Illumination Technology, National Taiwan University of Sci. & Tech., Taiwan (R.O.C.)
[2] Institute for Information Industry, Taiwan (R.O.C.)
Keywords: Luminance estimation, Deep learning, ResNet, Image analysis

ABSTRACT

This study used ResNet, a CNN-based deep learning model, to predict scene luminance levels from RGB images. ResNet effectively captures complex image features to accurately predict the scene luminance levels for both high-quality DSLR camera images and low-quality video images.

1 Introduction

Accurate prediction of scene luminance levels is essential for applications such as adaptive display technologies, autonomous vehicles, and intelligent surveillance systems. Traditional light sensors provide direct measurements but are limited in spatial coverage, the field of view and measuring directions of the sensors are different from the cameras, and are often expensive to deploy widely. To avoid the use of light sensors, this study explores the use of Convolutional Neural Networks (CNN) for a regression task aimed at predicting ambient light intensity from RGB images taken by high-end or low-end cameras.

Studies on evaluating scene lighting position, lighting direction, light distribution and color temperature [1][2][3] from image content are hot topics for image re-lighting and auto-white balancing. However, studies on evaluating scene luminance from image content are still very rare. The reason may be that it is very difficult to achieve accurate predictions. For example, when the illuminances of an outdoor scene changed from 10,000 lx to 100,000 lx, through the automatic exposure of the camera, there may be no difference in the tone and shadow of the photos. However, under widely varying weather conditions, such as sunny days and cloudy days, the human eyes can still recognize the differences. Computer vision should also be able to infer luminance levels from the sky color, clouds, and the sharpness of shadows.

We initially experimented with models such as VGG16 [4], MLP, and Random Forest to predict the scene luminance levels. However, the results were not as effective as expected, leading us to explore the ResNet architecture [5], which provided superior feature extraction capabilities. CNNs are known for their ability to automatically learn hierarchical features from large datasets, which makes them particularly suitable for complex visual tasks. Our dataset includes high-quality images captured by a DSLR camera (Nikon D7200) and low-quality images captured by a build-in video camera of Epson BT-45C AR smart-glasses, providing a unique opportunity to analyze the impact of different camera characteristics — such as automatic exposure and white balance settings — on the quality of scene luminance estimations.

2 Methods

2.1 Data Collection

To build a robust regression model for predicting ambient light intensity, we utilized two distinct datasets captured by a DSLR camera (Nikon D7200) and a build-in video camera of Epson BT-45C AR smart-glasses. Each dataset includes images taken under four different lighting conditions: sunny (including morning, noon and afternoon), cloudy, overcast, and indoor environments (as shown in Fig.1). This comprehensive data ensures the model is exposed to a wide range of light intensities and lighting directions, helping it generalize across various real-world settings.

In this study, we used an exposure meter to measure the Exposure Value (denoted as EV) under ISO speed 100. The averaged luminance L_{avg} of the viewing field can be roughly estimated by Equation 1.

$$L_{avg} = 2^{EV_{100}-3} \quad (1)$$

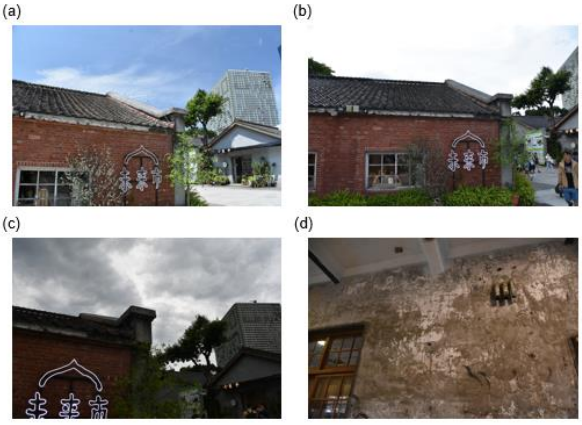

Fig. 1 Luminance levels (EVs) of a site in four lighting conditions:
(a) Sunny (EV: 11.6), (b) Cloudy (EV: 10.3),
(c) Overcast (EV: 7.5), (d) Indoor (EV: 4.7)

The DSLR image dataset provides high-resolution, 12-bit RAW images, which offer greater control over exposure and white balance, making them ideal for precise light measurement tasks. The high dynamic range (HDR) capabilities ensure the accurate capture of both shadowed and highlighted areas under different lighting conditions.

The video image data dataset consists of real-time, user-viewpoint images in 8-bit sRGB format, designed primarily for AR/VR applications. Although these images have lower resolution and dynamic range than the DSLR image dataset, they are essential for understanding how real-time performance affects scene luminance prediction in AR environments. Fig. 2 shows images captured simultaneously by the DSLR camera and the video camera at the same location. As can be seen, they differ in various aspects, including dynamic range, tone, shadow, and color.

(a)  (b) 

Fig. 2 Images captured simultaneously by two cameras: (a) DSLR camera, (b) video camera.

For each dataset, we rescaled the image dimensions to 224x224 pixels and applied image normalization to ensure uniform input for the model. This step was crucial to accommodate the differing resolutions and image formats between the two devices.

2.2 Model Selection

During the process of model selection, we evaluated several architectures, including VGG16, MLP, and Random Forest. However, these models did not perform as good as anticipated for the regression task. VGG16 [4], while being a deep convolutional network, lacks the residual connections that are critical for efficient learning in very deep networks. MLPs, although useful for lower-dimensional data, struggled with extracting meaningful features from high-dimensional image data. Similarly, Random Forest, despite its strengths with structured data, was less effective when applied to image-based regression tasks.

Ultimately, we selected the ResNet architecture [5] because of its use of residual blocks, which effectively address the vanishing gradient problem in deep networks. This allows the model to learn more complex and deep features from the data. Additionally, the pre-trained weights of ResNet enabled us to leverage powerful feature extraction without the need to train the network from scratch, making it particularly suitable for our task of predicting ambient light intensity from images. By modifying the final fully connected layer, we adapted ResNet from a classification task to a regression task, which allowed it to excel in predicting scene luminance levels from our dataset.

2.3 Model Training

We used the ResNet-18 model to estimate scene luminance level (EV values) from the collected image datasets and initialized the model with pre-trained weights from ImageNet to leverage its already well-developed feature extraction capabilities. This allowed us to focus on fine-tuning the model for the specific task of scene luminance level prediction, without needing to train the network from scratch.

The pre-trained ResNet-18 model was adapted by replacing the final fully connected layers used for classification with a single linear output layer, specifically designed for regression. This layer was tasked with predicting a continuous value representing the scene luminance levels (EVs) for each input image.

The datasets were split into training and testing sets using an 80/20 ratio, ensuring that each lighting condition, cloudy, overcast, sunny, and indoor, was represented in both sets. For the training process, we utilized the Mean Squared Error (MSE) loss function, which is well-suited for regression tasks aiming to minimize the difference between predicted and actual EVs. Additionally, the Adam optimizer was employed due to its ability to handle sparse gradients and noisy data, which are common when working with diverse image inputs from different sources.

2.4 Separate and Combined Training

The model was trained on three distinct configurations of the dataset:

DSLR images only: The model was trained exclusively on high-quality images from the DSLR camera (6,770 images) in Standard Mode to evaluate its ability to predict the scene luminance levels from precise, controlled data.

Video images only: The model was trained on the video camera dataset (approximately 13,000 images) to examine how well it could handle lower-quality, real-time images with automatic exposure and auto white balancing.

Combined dataset: Finally, the model was trained on a weighted mixture of the above mentioned two datasets (10,500 images) to assess its generalization capabilities when presented with high-quality and low-quality image data.

2.5 Evaluation and Testing

For each of the training configurations, we evaluated the model's performance on the test set by calculating the Mean Squared Error (MSE) and R^2 score. The results were then compared across the different datasets to determine how the model’s accuracy varied based on image source and quality.

3 Results

The performance of the ResNet model was evaluated across three datasets: (1) DSLR images only (Nikon D7200), (2) video images only (Epson BT 45C), and (3) combined dataset. The model's performance was assessed for each dataset in terms of Mean Squared Error (MSE), Mean Absolute Error (MAE), Loss, and Accuracy during the training, validation, and testing. The results are shown in **Table 1, Table 2,** and **Table 3**. The Accuracy represent the percentage of EV prediction errors less than ± 1 EV.

Table 1 Performance of the DSLR image dataset

Dataset	MSE	MAE	Loss	Accuracy
Training	0.05	0.175	0.025	99.4%
Validation	0.519	0.462	0.209	84.8%
Testing	0.624	0.532	0.263	83.6%

Table 2 Performance of the video image dataset

Dataset	MSE	MAE	Loss	Accuracy
Training	0.022	0.116	0.011	90.5%
Validation	0.412	0.406	0.182	87%
Testing	0.593	0.648	0.307	81.1%

Table 3 Performance of the combined dataset

Dataset	MSE	MAE	Loss	Accuracy
Training	0.047	0.169	0.023	99.7%
Validation	0.583	0.666	0.364	82.6%
Testing	0.718	0.718	0.391	72.7%

3.1 Performance of DSLR Image Dataset

The model achieved the highest performance on the DSLR image dataset during the training stage, with an MSE of 0.05, MAE of 0.175, Loss of 0.025, and a near-perfect Accuracy of 99.4%. However, a significant performance drop was observed during the validation and testing phases, with an MSE increasing to 0.519 and 0.624, respectively, and accuracy dropping to 83.6% in the testing phase. This drop suggests that while the model learns effectively from high-resolution, well-structured HDR images, it encounters challenges when exposed to unseen data, potentially due to overfitting even after data augmentation.

Fig.3 shows the predicted vs. actual EVs for the testing phase of this dataset. The model demonstrates a strong correlation between predictions and ground truth, with most data points aligning closely with the ideal diagonal line.

3.2 Performance of Video Image Dataset Performance

In the video image dataset, the model exhibited lower overall accuracy compared to the DSLR image dataset, but its performance was relatively more stable across the training, validation, and testing phases. In the training stage, the model achieved an MSE of 0.022, MAE of 0.116, Loss of 0.011, and an accuracy of 90.5%. While the accuracy dropped to 81.1% during testing, the overall performance degradation from training to testing was less severe compared to the DSLR image dataset, with the testing MSE increasing only to 0.593 from 0.412 in validation. This indicates that the model had greater generalization on the video camera dataset, despite the lower quality and resolution of the images.

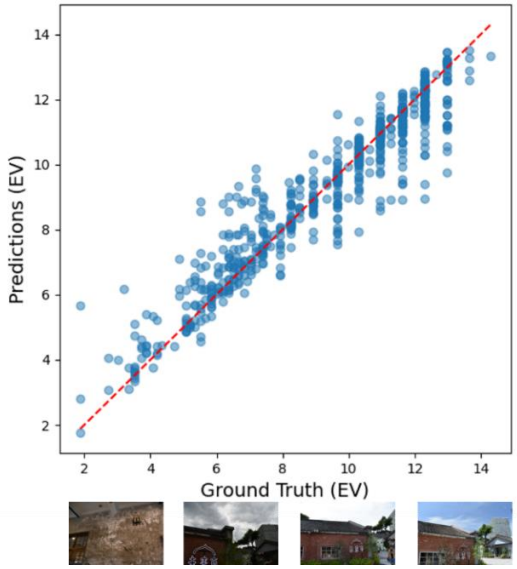

Fig. 3 Predicted and real luminance levels (EVs) of the DSLR image dataset

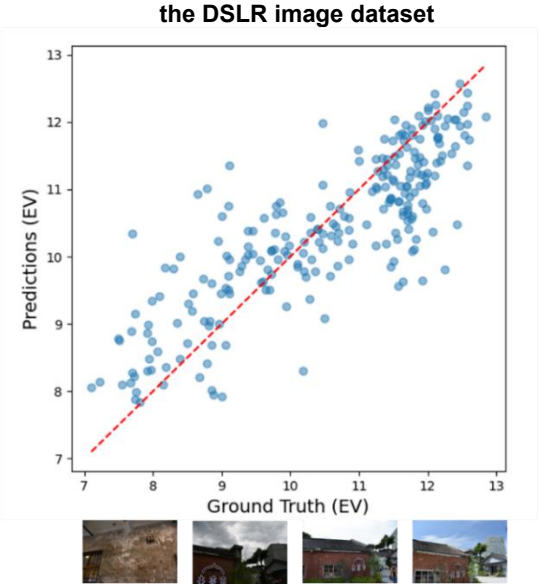

Fig. 4 Predicted and real luminance levels (EVs) of the video image dataset

In the video image dataset, the model performed well, though there was a noticeable drop in accuracy compared to the DSLR image dataset. As shown in Fig.4, the predicted EVs still closely follow the actual values. However, greater variations are observed, especially at the extremes: overestimation of EVs in low luminance scenes and underestimation in high luminance scenes.

3.3 Combined Dataset Performance

The combined dataset, which combines images from both the DSLR images and the video images, posed the greatest challenge for the model. During training, the model demonstrated excellent performance, achieving an MSE of 0.047, MAE of 0.169, Loss of 0.023, and an impressive accuracy of 99.7%. However, during the validation and testing phases, the performance significantly declined. The MSE increased to 0.583 in validation and 0.718 in testing, while accuracy dropped to 72.7% in testing. This result highlights the difficulties the model faces when attempting to generalize across datasets with differing image characteristics, such as resolution, dynamic range, and exposure settings.

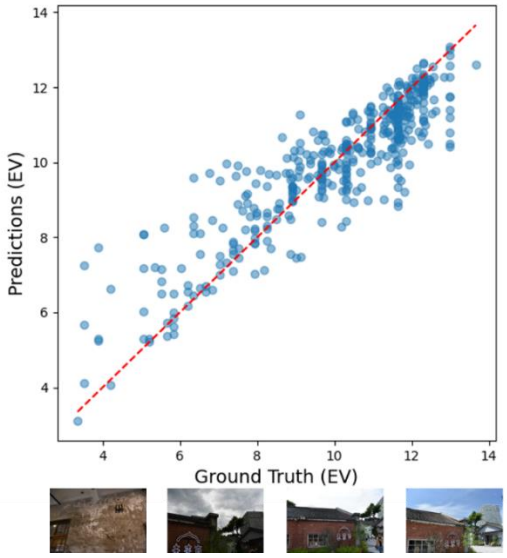

Fig. 5 Predicted and real luminance levels (EVs) for the combined dataset

The combined dataset is evidenced by the larger deviation between predicted and actual values shown as Fig.5. The model struggles more to generalize across both high-resolution (DSLR) and lower-resolution (video camera of the AR glasses) images, which is reflected in the wider spread of data points.

4 Discussion

The performance differences between the DSLR image and video image datasets can be largely attributed to variations in image quality and automatic adjustments. The DSLR image dataset benefited from the exposure, white balance, and image optimization features, such as contrast in detail enhancement. These factors provided the model with clearer and more consistent data, leading to higher accuracy. In contrast, the video image dataset faced challenges due to automatic adjustments in exposure and white balance, which introduced variability and reduced accuracy, especially in dynamic lighting conditions. The lower resolution and lack of image optimization in the video images further limited the model's ability to predict scene luminance levels as effectively as it did with the DSLR camera images.

When combining both datasets, the mixed dataset posed the greatest challenge, as the model had to generalize across heterogeneous data sources with varying levels of resolution, contrast, and dynamic range.

These findings suggest that future work may require specialized preprocessing techniques or domain adaptation methods to handle mixed-source data better and mitigate the negative impact of variability introduced by automatic settings, especially in lower-quality images.

5 Conclusions

This study demonstrated the effectiveness of using the ResNet model for predicting scene luminance levels (EVs) from both high-quality DSLR and low-quality video camera images under diverse lighting conditions. The model's ability to generalize across different image qualities, from high-resolution, controlled camera data to lower-quality, real-time images, makes it a promising solution for real-world applications. These include adaptive display systems, autonomous vehicles, and XR technologies, where dynamic ambient light conditions need to be accurately predicted. Future work will focus on enhancing the model's adaptability to even more varied datasets and improving its performance on lower-quality image sources.

6 Acknowledgements

This study is conducted under the "A Study on Environment-adaptive Color Conversion Technology Project" of the Institute for Information Industry which is subsidized by the Ministry of Economic Affairs of the Republic of China.

References

[1] E. Litvak and T. Kuflik, "Enhancing Cultural Heritage Outdoor Experience with Augmented-reality Smart Glasses," Personal and Ubiquitous Computing, vol.24, pp. 873-886, (2020).

[2] B.A.D. Marques, E.W.G. Clua, A.A.M. Montenegro and C.N. Vasconcelos, "Spatially and Color Consistent Environment Lighting Estimation using Deep Neural Networks for Mixed Reality," Computer & Graphics, vol. 102, pp. 257-268, (2021).

[3] S. Nathan and M.P. Beham, "LightNet: Deep Learning Based Illumination Estimation from Virtual Images." In: Bartoli, A., Fusiello, A. (eds) Computer Vision – ECCV 2020, vol 12537. Springer, (2020).

[4] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," ICLR 2015, (2015).

[5] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," CVPR, (2015).