INTRINSIC IMAGE EVALUATION ON SYNTHETIC COMPLEX SCENES

S. Beigpour[⋆]*, M. Serra*[⋆]*, J. van de Weijer*[⋆]*, R. Benavente*[⋆]*, M. Vanrell*[⋆]*, O. Penacchio*[⋆]*, D. Samaras*[†]

[⋆] Computer Vision Center, Universitat Autònoma de Barcelona, Spain
[†] Image Analysis Lab, Computer Science Dept., Stony Brook University, NY, USA

ABSTRACT

Scene decomposition into its illuminant, shading, and reflectance intrinsic images is an essential step for scene understanding. Collecting intrinsic image groundtruth data is a laborious task. The assumptions on which the ground-truth procedures are based limit their application to simple scenes with a single object taken in the absence of indirect lighting and interreflections. We investigate synthetic data for intrinsic image research since the extraction of ground truth is straightforward, and it allows for scenes in more realistic situations (e.g, multiple illuminants and interreflections). With this dataset we aim to motivate researchers to further explore intrinsic image decomposition in complex scenes.

***Index Terms*—** intrinsic images, synthetic data, reflectance modeling, illuminant estimation

1. INTRODUCTION

Decomposing the observed image into its intrinsic images by separating them into a shading, depth, reflectance, and illuminant color component is thought to be an essential step in scene understanding [1]. Most intrinsic image models have focused on separating reflectance from shading and are mainly based on the observation that reflectance changes produce sharp (high-frequency) intensity variations, while shading transitions are usually smoother (low-frequency) [2, 3]. Other methods [4, 5] have relaxed this assumption by adding cues on global reflectance sparsity. In [5], cues on color names were also used. Finally, a set of methods have differed from the classic intrinsic image decomposition approach. In [6], specularities and interreflections have been obtained assuming the dichromatic reflection model [7], while in [8, 9] the classic problem was pushed a step forward by estimating shape and illuminant information.

Initially, intrinsic image approaches showed interesting qualitative results on small sets of scenes [10, 11, 3]. It was not until Grosse *et al.* [12] built the MIT dataset for intrinsic image estimation that quantitative comparisons between methods could be done. This dataset has proved to be a useful evaluation tool. Its construction, however, presents some drawbacks that prevent any extension in terms of number of scenes or generalization to real scenes. Such a real ground truth collection is indeed very laborious: the same scene has to be captured twice, once with the original object and once after grey-painting the object, to obtain the shading ground truth; polarizing filters are used to separate specular from Lambertian reflectance; and interreflections need to be avoided because these would lead to false ground truth images [12]. As a result, the MIT dataset presents some drawbacks: only single object scenes are present, all of them are captured under white illumination, more complex and realistic lighting conditions (i.e. multiple illuminants) are not considered, and interreflections are absent. In [9], the MIT dataset was extended by synthetically relighting the images to obtain a multi-illuminant dataset. However, this has not solved the main drawback of the original dataset, namely the absence of complex realistic scenes with multiple objects.

Fig. 1. Two examples from the dataset under different illumination conditions. From top to bottom: white illuminant, single-colored light, and two distinct colored illuminants.

Therefore, evaluation of intrinsic image methods needs a new and more general dataset.

Obtaining a precise ground truth for complex real scenes, such as a landscape, would be hardly possible using the procedure described in [12]. Recently, the use of synthetic data to train and test complex computer vision tasks has attracted growing attention due to the increased accuracy with which 3D renderers visualize the world. In addition synthetic data allows for easy access to the ground truth, making it possible to prevent the expensive manual labeling process. Marin *et al.* [13] and Vazquez *et al.* [14] show that a pedestrian detector trained from virtual scenarios can obtain competitive results on real-world data. Liebelt and Schmid [15] use synthetic data to improve multi-view object class detection. Finally, Rodriguez *et al.* [16] generate synthetic license plates to train recognition system.

In this paper, we propose a synthetic dataset for intrinsic image evaluation which includes not only single object scenes with white illumination, but also multi-object scenes and multiple non-white illuminants with complex surrounding leading to interreflections (Figure 1). Multispectral sensors have been simulated in this paper in order to emulate a realistic visualization as described in [17, 18]. The objective of this new ground truth collection is to overcome the shortcomings of the existing datasets in intrinsic image evaluation and show an easy way to build ground truths for reflectance, shading, and illumination from synthetic data which allows the collection of a larger and more complex set of scenes. This dataset is available online[1] to further motivate research into more complex reflectance models. To validate our dataset, we evaluate and compare three existing methods [4, 5, 9].

2. MOTIVATION

Intrinsic image algorithms and datasets can be distinguished by their assumptions on the underlying reflectance models. Consider the reflection model [7] which models the color observation f^c with $c \in \{R, G, B\}$ as:

$$f^c(\mathbf{x}) = m(\mathbf{x}) \int_\omega s(\lambda, \mathbf{x})\, e(\lambda, \mathbf{x}) \rho^c(\lambda)\, d\lambda, \qquad (1)$$

where the integral is over all wavelengths λ of the visible spectrum ω. The material reflectance is given by $s(\lambda, \mathbf{x})$, $e(\lambda, \mathbf{x})$ is the spectrum of the illuminant, ρ^c is the camera sensitivity, and m is a scalar depending on the scene geometry (viewpoint, surface normal, and illuminant direction).

We will use this basic reflection model to demonstrate the differences between existing datasets and our dataset. In the MIT dataset [12] the illuminant is considered to be independent of $\mathbf{x}$ and white, i.e. $e(\lambda, \mathbf{x}) = 1$. This assumption is shared by most of the intrinsic image methods [8, 4, 5]. Recently, Barron and Malik [9] relaxed this assumption: they allowed the illuminant color to vary but only considered direct illumination (ignoring interreflections). Their assumption on the illuminant is given by $e(\lambda, \mathbf{x}) = e(\lambda, n(\mathbf{x}))$, where $n(\mathbf{x})$ is the surface normal at location $\mathbf{x}$. They construct a dataset by synthetically relighting the real-world MIT dataset [9].

In this paper, we go one step further and create a synthetic dataset by using rendering techniques from the computer graphics field. This allows us to remove the restriction other datasets put on $e(\lambda, \mathbf{x})$. The illuminant color and strength can change from location to location. This allows us to consider more complex reflection phenomena such as self-reflection and interreflection. To the best of our knowledge this is the first intrinsic image dataset which considers these more complex reflection models. In the next section we analyze rendering accuracy for such reflection phenomena.

Note that the above reflection model assumes that the materials have Lambertian reflectance. Even though specular materials can be accurately rendered, we exclude them from this dataset because most existing intrinsic image algorithms are not able to handle non-Lambertian materials. The MIT dataset [12] applies polarizing filters to provide both images with and without specular reflection.

3. SYNTHETIC INTRINSIC IMAGE DATASET

Recent advancements in digital 3D modeling programs have enabled the users to rely on these methods for graphical use, from digital animations and visual effects in movies to computer aided industrial design. Rendering is the process of generating a 2D image from a description of a 3D scene and is often done using computer programs by calculating the projection of the 3D scene model over the virtual image plane. Rendering programs are moving toward achieving more realistic results and better accuracy using physics-based models in optics. There are various softwares available which embed the known illumination and reflectance models [19].

In the current work, we have used Blender [20] to model the 3D scenes. YafaRay [21] is used as a rendering software for its photo-realism and physically plausible results. Both of these applications are free and open source.

3.1. Global Lighting for Scene Rendering

In order to obtain more photo-realistic lighting results for 3D scene rendering, a group of rendering algorithms have been developed which are referred to as global illumination. These methods, in addition to taking into account the light which reaches the object surface directly from a light source, called direct lighting, also calculate the energy which is reflected by other surfaces in the scene from the same light source. The latter is also known as indirect lighting. This indirect lighting is what causes the reflections, shadows, ambient lighting, and interreflections.

[1] http://www.cic.uab.cat/Datasets/synthetic_intrinsic_image_dataset

Fig. 2. Comparing different rendering methods: *direct lighting* (left) and *photon mapping* (right) on an example scene.

There are many popular algorithms for rendering global illumination (e.g, radiosity, raytracing, and image-based lighting). Among them, one of the most popular methods is a two pass method called photon mapping [22] developed by Henrik Wann Jensen. To achieve physically sound results and photo-realism in our dataset we make use of the photon mapping method embedded in YafaRay. Figure 2 shows the importance of indirect lighting. For this purpose we compare the final renderings of our dataset to the renderings which only consider direct lighting (one bounce). The former appears more realistic since diffuse interreflection is preserved.

3.2. Analysis of Color Rendering Accuracy

For synthetic datasets to be useful to train and evaluate computer vision algorithms, they should accurately model the physical reality of the real world. Therefore, in this section, we analyze the accuracy of color rendering based on the diagonal model as is typically done in graphics. To prevent propagating the full multispectral data, which is computationally very expensive, rendering engines approximate Eq. 1 with

$$\hat{f}^c = \int_\omega s\left(\lambda\right)\rho^c\left(\lambda\right)d\lambda \int_\omega e\left(\lambda\right)\rho^c\left(\lambda\right)d\lambda. \qquad (2)$$

Here we removed the dependence on $\mathbf{x}$ and the geometrical term m, and focus on the color content of f. In vector notation we could write this as

$$\hat{\mathbf{f}} = \mathbf{s} \circ \mathbf{e}, \qquad (3)$$

where we use bold to denote vectors, $\circ$ is the Hadamard product, and we replaced $\mathbf{s} = \int_\omega s\left(\lambda\right)\rho^c\left(\lambda\right)d\lambda$ and $\mathbf{e} = \int_\omega e\left(\lambda\right)\rho^c\left(\lambda\right)d\lambda$. In real scenes the light which is coming from objects is not only composed of reflection caused by direct lighting of the illuminant, but part of the light is reflected from other objects in the scene. Considering both direct and interreflection from another surface we can write:

$$\hat{\mathbf{f}} = \mathbf{s^1} \circ \mathbf{e} + \mathbf{s^2} \circ \mathbf{s^1} \circ \mathbf{e}, \qquad (4)$$

where the superscript is used to distinguish the material reflectance of different objects. The accuracy of the approximations in Eq. 3 and Eq. 4 is dependent on the shape and the number of sensors c considered. Typically rendering machines apply three sensors $c \in \{R, G, B\}$, however it is known that the rendering accuracy increases when considering more sensors [17, 18].

To test the accuracy of $\hat{f}^c$ we perform a statistical analysis. We use the 1269 Munsell color patches [23] and we compute both f^c and $\hat{f}^c$. For sensors ρ^c we use Gaussian shaped sensors which are equally spaced over the visible spectrum. We compare the reconstruction error $\varepsilon = \left\|\mathbf{f}\left(\mathbf{x}\right) - \hat{\mathbf{f}}\left(\mathbf{x}\right)\right\| \Big/ \left\|\mathbf{f}\left(\mathbf{x}\right)\right\|$ for the cases of three, six and nine sensors. We consider both single bounce (Eq. 3) and two bounces (Eq. 4). We use the standard D65 daylight illuminant. Dark patches were discarded because they cause the reconstruction error to be unstable.

	One bounce		**Two bounces**	
sensors	Mean (%)	Max (%)	Mean (%)	Max (%)
3	0.58	2.88	1.38	23.84
6	0.19	1.25	0.55	9.06
9	0.12	0.86	0.34	3.77

Table 2. Reconstruction error for single and two bounce reflection for 3, 6, and 9 sensors.

Table 2 shows the results of the experiment. For a single bounce the three sensor approximation, which is common in graphics, is acceptable and only leads to a maximum error of 2.88%. However, if we consider interreflections the maximum error reaches the unacceptable level of 23.84%. Based on these results, we have chosen to use a 6 sensors system to propagate the multispectral color information, resulting in a maximum error of 9.06%. This can be conveniently achieved by running existing rendering softwares (built for 3 channel propagation) twice for three channels [17, 18]. The final 6-D result image is projected back to a RGB image using linear regression. In the only available intrinsic image dataset for multi-illuminants [9], illuminants were introduced synthetically by using a 3 channel approximation, Based on our analysis, this is sufficent as it only considers direct lighting. Since this dataset only considers direct lighting, our analysis shows that this is sufficient. However, in the case of interreflections, synthetically relighting real-world scenes would introduce significant error.

Next, we address the importance of indirect lighting in scenes. For this purpose we compare the final renderings of our complex scenes to the renderings which only consider direct illumination (rendering programs allow for this separation). We compare the total energy in both renderings with the ratio $r = \sum_{\mathbf{x}} \left\|\mathbf{f}^1\left(\mathbf{x}\right)\right\| \Big/ \sum_{\mathbf{x}} \left\|\mathbf{f}^\infty\left(\mathbf{x}\right)\right\|$ where $\mathbf{f}^\infty$ is the final rendering and $\mathbf{f}^1$ is the single bounce rendering. For the

	Reflectance						Shading					
	Single Objects			Complex scenes			Single Objects			Complex scenes		
Method	WL	1L	2L	WL	1L	2L	WL	1L	2L	WL	1L	2L
Barron & Malik	0.082	0.099	0.102	0.020	0.059	0.039	0.043	0.046	0.054	0.011	0.014	0.014
Gehler *et al.*	0.089	0.113	0.123	0.018	0.067	0.040	0.043	0.045	0.051	0.007	0.009	0.009
Serra *et al.*	0.063	0.069	0.076	0.027	0.041	0.033	0.021	0.022	0.025	0.006	0.006	0.007

Table 1. LMSE results of three intrinsic image methods on our dataset. For clarity, errors for reflectance and shading are given separately. For both single objects and complex scenes, results for white illumination (WL), one illuminant (1L), and two illuminants (2L) are averaged.

nine complex scenes we found an average of $r = 0.83$, showing that a significant amount of lighting in the scene is coming from interreflections.

3.3. Proposed dataset

Our dataset consists of two set of images: single objects and complex scenes. In the first set, the aim is to simulate the work on MIT dataset. The second set is to our knowledge the first set of complex scenes for intrinsic image estimation which has an accurate ground truth, not only for the typical reflectance and shading decomposition, but also for the illuminant estimation. There are 8 objects in the first set. They vary in complexity for their shape and color distribution. The complex scenes, on the other hand, consist of various complex objects (e.g, furniture) which result in diffuse interreflections and complex shadows. Overall, there are 9 scenes in the second set. All the colors of the objects present in the scenes are taken from the Munsell colors since the multispectral reflectance values for them are recorded. Figure 3 shows examples of the ground truth we provide with the dataset. All the single object and complex scenes in our dataset are rendered under 4 different illumination conditions (i.e., white light, colored light, and 2 cases of multiple illuminants with distinct colors). This leads to a total of 32 images in the single-object set and 36 in the complex-scene set. The illuminants are randomly chosen from a list of Planckian and non-Planckian lights from the Barnard dataset [24].

Fig. 3. Two examples of ground-truth decomposition. From left to right: the rendered scene, reflectance component, and shading-illumination.

4. EXPERIMENTS

In order to show that our dataset is suitable for evaluating intrinsic image methods, we compare three different models for intrinsic image decomposition which are currently state-of-the-art [9, 4, 5]. For this experiment, we have used the publicly available codes of the methods, with the default parameters. Therefore, we have not trained the models on this specific dataset.

For each of the subsets of our dataset, namely single objects and complex scenes, we have analyzed the three methods on three illumination conditions: white light (WL), one non-white illuminant (1L), and two non-white illuminants (2L). The mean results for each illumination condition have been computed.

Errors have been evaluated by using the local mean squared error (LMSE) and considering the three RGB channels of the color image [12]. As reflectance images can be recovered only up to a scale factor, we have multiplied the estimated reflectance images by an α factor which has been fitted for each local patch to minimize the MSE.

Table 1 summarizes the results obtained for reflectance and shading. As expected, the error for all methods increases when the illuminant is not white. The shading evaluation is relatively invariant to illuminant changes because it discards color information. The lower errors on the complex scenes are caused by large uniform colored objects which result in low LMSE. The method of Serra et al. [5] obtained the best results. However, visual inspection of the results revealed that the design of new error measures is a necessity for intrinsic image evaluation, i.e. visual ranking of the accuracy did often not agree with the LMSE error based ranking.

5. CONCLUSIONS

This paper shows that synthetic data constitute a valid medium for intrinsic image evaluation. It encourages the collection of large intrinsic synthetic image datasets which allow evaluation also in complex scenes under multiple illuminants.

Acknowledgements: This work has been supported by projects TIN2009-14173 and TIN2010-21771-C02-1 of Spanish Ministry of Science and Innovation, and 2009-SGR-669 of Generalitat de Catalunya. Dimitris Samaras acknowledges partial support from the Subsample Project of the DIGITEO Institute, France, and from a gift by Adobe Corp. We acknowledge Peter Gehler and Martin Kiefel for kindly providing the results of their algorithm on our data base.

6. REFERENCES

[1] H.G. Barrow and J.M. Tenenbaum, "Recovering intrinsic scene characteristics from images," *Computer Vision Systems*, pp. 3–26, 1978.

[2] E.H. Land, "The retinex theory of colour vision," *Scientific American*, vol. 237, no. 6, pp. 108–129, 1977.

[3] M.F. Tappen, W.T. Freeman, and E.H. Adelson, "Recovering intrinsic images from a single image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 9, pp. 1459–1472, 2005.

[4] P.V. Gehler, C. Rother, M. Kiefel, L. Zhang, and B. Schölkopf, "Recovering intrinsic images with a global sparsity prior on reflectance," in *Advances in Neural Information Processing Systems (NIPS)*, 2011, pp. 765–773.

[5] M. Serra, O. Penacchio, R. Benavente, and M. Vanrell, "Names and shades of color for intrinsic image estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 278–285.

[6] S. Beigpour and J. van de Weijer, "Object recoloring based on intrinsic image estimation," in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 327–334.

[7] S.A. Shafer, "Using color to separate reflection components," *Color Research and Application*, vol. 10, no. 4, pp. 210–218, 1985.

[8] J.T. Barron and J. Malik, "High-frequency shape and albedo from shading using natural image statistics," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 2521–2528.

[9] J.T. Barron and J. Malik, "Color constancy, intrinsic images, and shape estimation," in *European Conference on Computer Vision (ECCV)*, 2012, pp. 55–70.

[10] Y. Weiss, "Deriving intrinsic images from image sequences," in *International Conference on Computer Vision*, 2001, pp. 68–75.

[11] G. Finlayson, M. Drew, and C. Lu, "Intrinsic images by entropy minimization," in *European Conference on Computer Vision*, 2004, pp. 582–595.

[12] R. Grosse, M.K. Johnson, E.H. Adelson, and W.T. Freeman, "Ground truth dataset and baseline evaluations for intrinsic image algorithms," in *IEEE International Conference on Computer Vision*, 2009, pp. 2335–2342.

[13] J. Marin, D. Vázquez, D. Gerónimo, and A.M. López, "Learning appearance in virtual scenarios for pedestrian detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 137–144.

[14] D. Vázquez, A.M. López, and D. Ponsa, "Unsupervised domain adaptation of virtual and real worlds for pedestrian detection," in *International Conference on Pattern Recognition*, 2012.

[15] J. Liebelt and C. Schmid, "Multi-view object class detection with a 3d geometric model," in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 1688–1695.

[16] J. Rodriguez-Serrano, H. Sandhawalia, R. Bala, F. Perronnin, and C. Saunders, "Data-driven vehicle identification by image matching," in *Computer Vision–ECCV 2012. Workshops and Demonstrations*. Springer, 2012, pp. 536–545.

[17] M.S. Drew and G.D. Finlayson, "Multispectral processing without spectra," *Journal of the Optical Society of America A*, vol. 20, no. 7, pp. 1181–1193, 2003.

[18] B. Darling, J.A. Fewerda, R.S. Berns, and T. Chen, "Real-time multispectral rendering with complex illumination," in *19th Color and Imaging Conference*, 2010, pp. 345–351.

[19] M. Pharr and G. Humphreys, *Physically Based Rendering: From Theory to Implementation*, The Morgan Kaufmann series in interactive 3D technology. Elsevier Science, 2010.

[20] "http://www.blender.org," .

[21] "http://www.yafaray.org," .

[22] H.W. Jensen, *Realistic image synthesis using photon mapping*, A.K. Peters, Ltd., Natick, MA, USA, 2001.

[23] "http://www.uef.fi/spectral/spectral-database," Last accessed on May 28, 2013.

[24] K. Barnard, L. Martin, B. Funt, and A. Coath, "A data set for color research," *Color Research & Application*, vol. 27, no. 3, pp. 147–151, 2002.