



# Modeling, estimation and evaluation of intrinsic images considering color information

A dissertation submitted by **Marc Serra Vidal**  
at Universitat Autònoma de Barcelona to fulfil  
the degree of **Doctor en Informàtica**.

Bellaterra, June 2015

Directors: **Dr. Robert Benavente**  
Centre de Visió per Computador  
Dept. de Ciències de la Computació, Universitat Autònoma de Barcelona

**Dr. Olivier Penacchio**  
School of Psychology and Neuroscience  
University of St Andrews



---

This document was typeset by the author using L<sup>A</sup>T<sub>E</sub>X 2<sub>ε</sub>.

The research described in this book was carried out at the Computer Vision Center, Universitat Autònoma de Barcelona.

Copyright © 2015 by Marc Serra Vidal. All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the author.

ISBN 978-84-943427-4-5

To my family and friends



# Agraïments

Vull començar dedicant aquest treball als meus pares. Ells em van animar a emprendre aquesta aventura ara fa uns anys, quan jo no ho veia gaire clar. Tots dos han estat sempre un exemple a seguir i si he arribat fins aquí és gràcies a ells, que sempre han confiat en mi i m'han ofert un suport incondicional en tots els moments difícils.

Estic agraït a la Maria Vanrell, que em va oferir ara fa uns quants anys la possibilitat d'incorporar-me al grup de recerca de color del Centre de Visió per Computador i de fer aquest doctorat, codirigit pel Robert Benavente i l'Olivier Penacchio. També agraeixo a la Maria les múltiples idees que ha aportat al meu treball i l'oportunitat que em va oferir de fer una estada de recerca a Nova York amb el Dimitris Samaras.

Als meus directors, el Robert Benavente i l'Olivier Penacchio, els agraeixo tota l'ajuda rebuda durant aquests anys i la paciència que han tingut amb mi. Vull agrair a l'Olivier el tracte personal que he rebut durant tots aquests anys i la seva implicació en aquest treball, fins i tot quan uns quants milers de quilòmetres de distància s'han interposat entre nosaltres.

I wish to express my sincere thanks to Dimitris Samaras, who accepted me in his research group at Stony Brook University for some months. I have excellent memories of the time I spent there, especially of the informal meetings we had on the train or while we were having dinner. Dimitris has been both a director and a friend. I really appreciate his personal treatment, the confidence he has put in me, and the multiple good ideas he has provided during all these years. I also take this opportunity to express gratitude to my lab mates at SBU, Kota Yamaguchi, Kiwon Yun, Vicente Ordonez, Yifan Peng, Xufeng Han and Tomas Yago, who made me feel at home and became my friends.

A la gent del grup de recerca els agraeixo tots els consells, tant acadèmics com personals, que m'han proporcionat durant aquests anys. Ha estat un plaer col·laborar amb el Joost van de Weijer i la Shida Beigpour. Agraeixo al Ramon Baldrich que em deixés ocupar el seu despatx de la ETSE per muntar-hi un laboratori clandestí d'adquisició d'imatges. Als estudiants de doctorat del grup, la Camp Davesa, l'Ivet Rafegas i el Ricard Balagué els hi desitjo tota la sort del món.

Fora del centre de recerca, vull agrair a la Shannon Silvestre, la Jessica Arribas, l'Anna Casablanca, la Yainuvis Socarrás, la Mercè Ortega-Villaizán i la Dolors Giralt tots els bons moments que hem passat junts i als meus companys de pis de Cerdanyola, Albert Forniés, Núria Foguet i Cristina Llecha que m'hagin aguantat i m'hagin fet sentir com a casa durant els primers anys del doctorat.

També vull agrair als meus companys del SAF, que han acabat sent amics dins

i fora del gimnàs, que m'hagin ajudat, possiblement sense saber-ho, a mantenir l'equil·libri emocional durant aquest últim any de doctorat. D'aquest curs amb la Gemma Prats, la Laura López, el Xavi Raba, la Sònia Cobos, l'Anna Font, el Daniel Gimenez, l'Izar Capel, la Lidia Medina, el Sergi Delgado i la Naroa Uría sempre en guardaré un gran record.

Finalment, vull fer un esment especial a la Patricia Márquez, amiga des de fa molts anys i també companya de penes i al·legries. Ella ha estat el suport moral més gran que he tingut durant tots aquests anys. Poca gent em coneix tant bé com ella. Pat, gràcies per ser com ets!

# Resum

Els valors dels píxels de les imatges són el resultat d'una combinació d'informacions visuals provinents de múltiples fonts. Recuperar la informació dels múltiples factors que han produït una imatge sembla un problema molt difícil. Tanmateix, els éssers humans desenvolupem l'habilitat d'interpretar les imatges i podem reconèixer i aïllar propietats físiques de l'escena.

Les imatges que descriuen una sola característica física d'una escena s'anomenen *imatges intrínseques*. Aquestes imatges simplificarien la majoria de processos de la visió per computador, que sovint es veuen afectats pels diversos efectes que normalment trobem en les imatges naturals (ombres, especularitats, interreflexions, etc.) En aquesta tesi analitzem el problema de l'estimació d'imatges intrínseques des de diferents punts de vista, com per exemple la formulació teòrica del problema, les cues visuals que ens ajuden a estimar certes propietats intrínseques de les imatges o els mecanismes d'avaluació del problema.

Primer introduïm breument l'origen del problema de la descomposició d'imatges intrínseques i també parlem del context del problema i d'alguns temes que hi estan relacionats. Llavors, presentem una revisió exhaustiva de la bibliografia d'imatges intrínseques en el camp de la visió per computador, proporcionant una descripció detallada i organitzada de les tècniques per a l'estimació d'imatges intrínseques que han aparegut fins ara. En aquesta revisió analitzem com algunes assumpcions habituals sobre les escenes han afectat la formulació del problema. També estudiem com algunes cues d'informació, basades en regularitats sobre les escenes i les imatges, s'han utilitzat per estimar imatges intrínseques. D'altra banda, també examinem els mecanismes d'avaluació d'imatges intrínseques existents, estudiant les bases de dades i les mètriques actuals. A més a més, analitzem l'evolució d'aquest camp de recerca i n'identifiquem les tendències actuals.

Sovint, en el camp de la visió per computador, la informació del color ha estat ignorada. Tanmateix, tal i com es pot veure en la nostra revisió del problema, el color és molt útil en l'estimació d'imatges intrínseques. En aquest treball presentem un mètode de descomposició d'imatges intrínseques que utilitza dos atributs de color diferents que es combinen en un marc probabilístic. El primer està basat en la descripció semàntica del color que fan servir els humans i proporciona una descripció dispersa de reflectàncies en una imatge. L'altre es basa en un anàlisi de les distribucions de color, que connecta els màxims locals dins de l'histograma de color de la imatge. Aquest atribut proporciona una descripció consistent de superfícies que comparteixen la mateixa reflectància i aporta estabilitat als noms de colors en regions de

la imatge afectades per ombres i també en regions pròximes a especularitats.

D'altra banda, la majoria dels mètodes de descomposició d'imatges intrínseques fins ara han assumit que les escenes estan il·luminades per una "llum blanca" i han ignorat completament els efectes dels sensors de la càmera en les imatges. Tots dos factors, però, afecten els valors de les imatges resultants durant el procés d'adquisició. En aquest treball analitzem la formulació teòrica del problema de descomposició d'imatges intrínseques i proposem un nou marc, més general, on es modelitzen els efectes tant dels sensors de la càmera com del color de l'il·luminant. En aquesta nova formulació hi introduïm un nou component, anomenat reflectància absoluta, que és invariant a tots dos efectes. A més a més, demostrem que qualsevol coneixement sobre el color de l'il·luminant o sobre els sensors de la càmera es pot utilitzar per millorar les reflectàncies estimades dels diferents mètodes de descomposició d'imatges intrínseques. També mostrem que els mètodes existents, que normalment ignoren el color de l'il·luminant i els sensors de la càmera, inclouen errors molt grans en les seves reflectàncies estimades.

Finalment, analitzem els mecanismes d'avaluació d'imatges intrínseques, que han evolucionat constantment durant aquesta última dècada. Tot i que s'han presentat diverses bases de dades, la seva construcció és un problema complicat i totes presenten múltiples inconvenients, com el nombre reduït i la poca diversitat d'imatges o la falta d'informació sobre determinades propietats intrínseques de les escenes (la profunditat o orientació dels objectes que apareixen a les imatges, el color i direcció de l'il·luminant, etc.). En aquesta tesi presentem dues bases de dades per a l'avaluació d'imatges intrínseques. Una és una base de dades calibrada que inclou informació sobre l'il·luminant de l'escena i els sensors de la càmera. Aquesta base de dades s'ha utilitzat per validar experimentalment el marc teòric per a la descomposició d'imatges intrínseques presentat en aquesta tesi. La segona base de dades s'ha construït mitjançant tècniques de gràfics per computador i conté imatges, tant d'objectes simples com d'escenes complexes, adquirides amb diferents condicions d'il·luminació. En aquest treball demostrem que amb programari de gràfics per computador i motors de representació gràfica, és possible construir bases de dades molt grans i realistes per a l'avaluació d'imatges intrínseques.



# Abstract

Image values are the result of a combination of visual information coming from multiple sources. Recovering information from the multiple factors that produced an image seems a hard and ill-posed problem. However, it is important to observe that humans develop the ability to interpret images and recognize and isolate specific physical properties of the scene.

Images describing a single physical characteristic of an scene are called intrinsic images. These images would benefit most computer vision tasks which are often affected by the multiple complex effects that are usually found in natural images (*e.g.* cast shadows, specularities, interreflections...).

In this thesis we will analyze the problem of intrinsic image estimation from different perspectives, including the theoretical formulation of the problem, the visual cues that can be used to estimate the intrinsic components and the evaluation mechanisms of the problem.

We first give a brief introduction on the background and the nature of the problem of intrinsic image estimation and some of its closely related topics. Then, we present an exhaustive review of the literature of intrinsic images in the field of computer vision, giving a comprehensive and organized description of the existing techniques for intrinsic image estimation. In our review we analyze how common simplifying assumptions about the world have modified the formulation of the problem of intrinsic image decomposition and also how different information cues based on regularities about the scenes and images have been used to estimate intrinsic images. We also examine the evaluation mechanisms that have been used so far in this problem. We analyze the existing databases and metrics, discuss the evolution of the problem and identify the recent trends in the field.

Color information has been frequently ignored in the field of computer vision. However, as it can be seen in our review, color has proved to be extremely useful in the estimation of intrinsic images. In this work we present a method for intrinsic image decomposition which estimates the intrinsic reflectance and shading components of a single input image using observations from two different color attributes combined in a probabilistic framework. One of them, based on the semantic description of color used by humans, provides a sparse description of reflectances in an image. The other, based on an analysis of color distributions in the histogram space which connects local maxima, gives us a consistent description of surfaces sharing the same reflectance, providing stability of color-names in shadowed or near highlight regions of the image.

Moreover, most methods for intrinsic image decomposition have usually assumed

“white light” in the scenes and have completely ignored the effect of camera sensors in images. However, both factors strongly influence the resulting image values during the acquisition process. In this work we analyze the theoretical formulation underlying the decomposition problem and propose a generalized framework where we model the effects of both the camera sensors and the color of the illuminant. In this novel formulation we introduce a new reflectance component, called absolute reflectance, which is invariant to both effects.

Furthermore, we demonstrate that any knowledge of the color of the illuminant or the camera sensors can be used to improve the reflectance estimates of different existing methods for intrinsic image decomposition. We also show that existing methods, which usually ignore the color of the illuminant and the camera sensors, include large errors in their reflectance estimates.

Finally, we analyze the evaluation mechanisms of intrinsic images, which have continuously evolved during the last decade. Although multiple datasets have been presented, building these datasets has proved to be a challenging problem in itself and current ground truth collections present multiple drawbacks, such as the small number and diversity of scenes or the lack of ground truth information for specific intrinsic components (the depth or surface orientation of the objects in the image, the color and direction of the illuminant, etc.). In this thesis we present two datasets for intrinsic image evaluation. One is a calibrated dataset which includes ground truth information about the illuminant of the scene and the camera sensors. This dataset is used in this work to experimentally validate the theoretical framework for intrinsic image decomposition proposed in this thesis. The second dataset uses synthetic data and contains both simple objects and complex scenes under different illumination conditions. In this work we demonstrate that it is possible to build large and realistic datasets for intrinsic image evaluation using computer graphics software and rendering engines.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Visual Descriptions of Scenes . . . . .	2
1.2	Outline of the Human Visual System . . . . .	5
1.3	Background on Human Visual Perception . . . . .	7
1.4	Intrinsic Images in Computer Vision . . . . .	11
1.4.1	Color Constancy . . . . .	14
1.4.2	Sensor Calibration . . . . .	15
1.4.3	Intrinsic Image Evaluation . . . . .	15
1.5	Scope of the thesis . . . . .	15
<b>2</b>	<b>Review on Computational Intrinsic Images</b>	<b>17</b>
2.1	Problem formulation and simplifying assumptions about the scenes . .	19
2.1.1	Discussion . . . . .	20
2.2	Visual cues based on scene and image regularities . . . . .	20
2.2.1	Shading Smoothness . . . . .	20
2.2.2	Texture Structure . . . . .	22
2.2.3	Color Sparsity . . . . .	23
2.2.4	Discussion . . . . .	23
2.3	Intrinsic image decomposition techniques . . . . .	25
2.3.1	Classification of image derivatives . . . . .	25
2.3.2	Learning-based approaches . . . . .	26
2.3.3	Energy functions optimization . . . . .	26
2.3.4	Discussion . . . . .	30
2.4	Intrinsic Image Evaluation . . . . .	32
2.4.1	Qualitative examples . . . . .	32
2.4.2	Datasets . . . . .	33
2.4.3	Metrics . . . . .	37
2.4.4	Discussion . . . . .	40
2.5	Conclusion and Perspectives . . . . .	41
<b>3</b>	<b>Shades and Names of Color for Intrinsic Image Estimation</b>	<b>45</b>
3.1	Motivation . . . . .	46
3.2	Our approach . . . . .	46
3.2.1	Color-name descriptor . . . . .	47

3.2.2	Color-shade descriptor . . . . .	48
3.2.3	Method outline . . . . .	49
3.3	Reflectance recovery using MRF inference . . . . .	50
3.3.1	Singleton potential: color name . . . . .	51
3.3.2	Pairwise potential: color shade . . . . .	51
3.3.3	MRF output . . . . .	52
3.4	Adding global scene coherence . . . . .	52
3.5	Experiments . . . . .	54
3.6	Discussion and conclusion . . . . .	59
<b>4</b>	<b>A General Framework Based on the Photometry of Intrinsic Images</b>	<b>65</b>
4.1	Motivation . . . . .	65
4.2	Reflectance and color fundamentals . . . . .	67
4.2.1	Color image formation: physics . . . . .	67
4.2.2	Color image formation: sensors . . . . .	69
4.2.3	Joint illumination/sensor modeling for intrinsic images . . . . .	70
4.3	A general model for intrinsic image estimation . . . . .	71
4.3.1	Model particularities and relation to previous models . . . . .	72
4.4	Validation Experiments . . . . .	77
4.4.1	Experiment 1: Synthetic data . . . . .	78
4.4.2	Experiment 2: Natural images . . . . .	80
4.4.3	Experiment 3: Laboratory-acquired images . . . . .	81
4.5	Discussion and conclusion . . . . .	87
<b>5</b>	<b>New Datasets for Intrinsic Image Evaluation</b>	<b>89</b>
5.1	A Calibrated Dataset for Intrinsic Image Estimation . . . . .	90
5.1.1	Methods and Materials . . . . .	90
5.1.2	Acquisition Process . . . . .	91
5.1.3	Experiments . . . . .	94
5.2	Synthetic Intrinsic Image Dataset . . . . .	97
5.2.1	Motivation . . . . .	97
5.2.2	Global Lighting for Scene Rendering . . . . .	98
5.2.3	Analysis of Color Rendering Accuracy . . . . .	98
5.2.4	Proposed dataset . . . . .	101
5.2.5	Experiments . . . . .	102
5.3	Discussion and Conclusion . . . . .	105
<b>6</b>	<b>Conclusions and Future Work</b>	<b>107</b>
6.1	Conclusions . . . . .	107
6.2	Future Work . . . . .	109
6.2.1	Future Research and Applications . . . . .	110
6.2.2	Recent Trends in the Field of Intrinsic Images . . . . .	111
<b>A</b>	<b>Discretization of the color-name descriptor</b>	<b>113</b>
<b>B</b>	<b>Illumination Conditions</b>	<b>115</b>

*CONTENTS*

ix

<b>C Database images</b>	<b>117</b>
C.1 Calibrated Dataset . . . . .	117
C.2 Synthetic Dataset . . . . .	125
<b>Bibliography</b>	<b>129</b>
<b>List of Publications</b>	<b>141</b>



# List of Tables

2.1	Classification of the methods based on the simplifying assumptions they make about the scenes. † Specularities are included in the shading image. * Grayscale images are used. . . . .	21
2.2	Classification of the methods according to their assumptions about the most common regularities in the scenes. . . . .	24
2.3	Classification of the methods based on their decomposition technique as well as their inputs. The outputs of the methods have also been included in this table for the sake of completion. Other outputs: (1) Isolated shadow information. (2) Specular component. (3) Isolated noise effects. (4) Model of illumination. (5) Direct and indirect irradiance information. (6) Object and attribute labels. (7) Texture component. (8) Optical Flow. * Grey-scale reflectance images. † Color shading images. . . . .	31
2.4	Classification of the methods based on the evaluation tools that the authors have used. Databases: (1) Psychophysics [55].(2) Yale Face [61]. (3) Crayola [145]. (4) BOLD [90]. (5) MIT - 16 objects [71]. (6) MIT - 20 objects [71]. (7) extended MIT [20]. (8) NYU Depth [139]. (9) MPI-Sintel [35]. (10) images from Flickr. (11) Intrinsic Images in the Wild [28]. Metrics: (a) MSE (Eq. 2.4). (b) LMSE (Eq. 2.5). (c) aLMSE (Eq. 2.6). (d) Corr (Eq. 2.7). (e) LCorr (Eq. 2.8). (f) Average Rank. (g) DSSIM (Eq. 2.11). (h) MAE (Eq. 2.14). (i) WHDR (Eq. 2.12). * Only results for 13 objects are provided. . . . .	43
2.5	Comparison of the existing datasets for intrinsic image evaluation. * Sequences with different numbers of frames (50 on average). . . . .	44
3.1	Results on the MIT-16 dataset (16 objects) with different error metrics. Shen-SRC results are computed on a subset of 13 objects ('deer', 'squirrel' and 'dinosaur' results were not available). . . . .	57
3.2	Results on the MIT-20 dataset (20 objects) with different error metrics. . . . .	57
3.3	Shading and reflectance images recovered by previous algorithms and our approach from images of the MIT database. Values below each decomposition are the corresponding correlation measures. . . . .	58
3.4	MSE and correlation results on the MIT-16 and MIT-20 datasets. . . . .	59
3.5	LMSE and aLMSE results on the MIT-16 and MIT-20 datasets. . . . .	60

3.6	Qualitative results for the “animal” images in the MIT dataset. . . . .	61
3.7	Qualitative results for the “printed paper” images in the MIT dataset.	61
3.8	Qualitative results for the “painted object” images in the MIT dataset.	62
3.9	Qualitative results for the 4 extra images in the MIT-20 dataset. . . . .	63
4.1	Specific formulations of the model resulting from common assumptions.	75
4.2	Summary of how previous methods for intrinsic image decomposition are related to our general model. . . . .	76
4.3	Description of the scenarios used in the experiments. . . . .	78
4.4	Numerical results (angular error in degrees) of the experiments showed in Figure 4.3. . . . .	80
5.1	LMSE results of reflectance image estimates in different scenarios. . . . .	95
5.2	Reconstruction error for single and two bounce reflection for 3, 6, and 9 sensors. . . . .	100
5.3	LMSE results of three intrinsic image methods on the single object scenes included in our dataset. Errors for reflectance and shading are given separately for the sake of clarity. Results for white illumination (WL), one illuminant (1L), and two illuminants (2L) are averaged. . . . .	104
5.4	LMSE results of three intrinsic image methods on the complex scenes included in our dataset. Errors for reflectance and shading are given separately for the sake of clarity. Results for white illumination (WL), one illuminant (1L), and two illuminants (2L) are averaged. . . . .	104
5.5	Comparison of the existing datasets for intrinsic image evaluation. * Sequences with different numbers of frames (50 on average). . . . .	106



# List of Figures

1.1	Examples of techniques used to represent depth before the Renaissance period. (a) A cave painting found in Valltorta, Valencia (Spain), where relative position has been used to represent depth. (b) Transparent materials were represented in ancient Egypt in the painting <i>The goddess Hathor welcomes Sethos I</i> . Occlusion was also used in this painting to represent depth. (c) An example of ancient Greek art where occlusion and relative position are used to convey a sense of depth. (d) A sample of Byzantine art, the altarpiece <i>Maestà of Duccio</i> , where occlusion, scale and relative position are used. . . . .	3
1.2	Examples of techniques that came out during the Renaissance period. Leonardo da Vinci's painting <i>L'Ultima Cena</i> (a) is a clear example of linear perspective. Sfumato was used in one of the most famous art pieces, Leonardo da Vinci's painting <i>La Gioconda</i> (b). Giovanni Baglione's painting <i>Amor sacro e amor profano</i> (c) provides a great example of chiaroscuro. These techniques convey a vivid sensation of shape and volume. . . . .	4
1.3	Image acquisition process. The light source, the materials of the objects and the response of the camera sensors influence the resulting image values. Afterwards, during the quantization process, geometric and colorimetric discretizations are applied to the image. . . . .	5
1.4	Response of the different photoreceptor cells in the retina of human beings with normal color vision. R indicates rods. S,M and L indicate the different types of cones. . . . .	6
1.5	Examples of brightness induction (a) and color constancy (b). In (a), although both grey squares have exactly the same color, the square surrounded by black is perceived brighter than the square surrounded by white, a phenomenon referred to as brightness induction. In (b), we see the same scene under two different illumination conditions: A reddish light in the image on the left and a bluish light in the image on the right (data from [17]). Although these illumination conditions result in different pixel values, we perceive the same colors in both images. This phenomenon is referred to as color constancy. . . . .	7

1.6	Representation of the ratio-invariance property at edge intersections described in [64]. An illumination edge (the vertical one) dividing regions A and C from regions B and D crosses a reflectance edge (the horizontal one) separating regions A and B from regions C and D) and the ratio-invariance property, $\frac{A}{B} = \frac{C}{D}$ , holds. . . . .	9
1.7	The perception of transparency [120]. The transparency perceived in (a) disappears when the figural unity is broken (b) or when certain color conditions are not met (c). . . . .	9
1.8	The influence of 3D interpretation on brightness and lightness perception (image adapted from [10]). The edges in (a) would be equally classified by local methods such as the Retinex [104], although the upper one is a result of a variation in illumination (b), while the lower one is due to a change in reflectance (c). . . . .	10
1.9	Image adapted from [11]. Different illumination, shape and reflectance conditions result in the same image. Are some of these combinations more likely than others? . . . . .	11
1.10	Computational model for the recovery of intrinsic image presented in the work of Barrow and Tenenbaum [23]. In this model, an input intensity image (first layer) is decomposed into multiple intrinsic images (subsequent layers) using intra-image continuity (circles), inter-image constraints (vertical lines) and further local processes (X marks) in each layer to inhibit continuity constraints. . . . .	12
2.1	Intrinsic components as illustrated by Barrow and Tenenbaum in [23] are shown in the top row, while actual representations of these intrinsic components can be found in the bottom row. Representations for depth and orientation in [23] are based on the 2.5D sketch of Marr [117]. . .	18
2.2	Input images for different methods. (a) Image containing cast shadows [49]. (b) Image with specularities [143]. (c) Highly textured images [137].	33
2.3	Graffiti image from [27]. (a) Graffiti painted on a wall. (b) Graffiti covered with white paint. . . . .	34
2.4	Sample image from the Crayola database [145]. (a) Original image: wrinkled paper colored with a Crayola marker. (b) Shading image: the Crayola marker is invisible in the Green channel of the image. (c) Reflectance image: result of dividing the original image by the shading image. . . . .	34
2.5	Images in the MIT dataset [71]. (a) The 16 original objects that composed the dataset. (b) The 4 extra objects that can be found on the authors' website. . . . .	35
2.6	Barron and Malik extended the MIT dataset [71]. Example of extensions provided for the sun object. (a) Estimating shape and illuminant information [20]. (b) Adding complex illumination models [19]. . . . .	36
2.7	Synthetic image of a doll in [32]. (a) Original image. (b) Shading image. (c) Reflectance image. . . . .	37
2.8	Example of MPI-Sintel dataset groundtruth [35]. (a) Original image. (b) Normal map. (c) Shading image. (d) Reflectance image. . . . .	37

2.9 Example to illustrate the behaviour of different error metrics, given an original image (a), its ground truth reflectance (b) and shading (c) and the computed estimates for reflectance (d) and shading (e) using a Retinex-like method which misclassifies an image derivative. . . . . 38

3.1 Shading and reflectance images recovered with our method. . . . . 45

3.2 Color-name descriptor. (a) Plotted volumes represent those values in the RGB color space with probability equal to 1 of being one of the 11 universal colors according to [30]. The space between the volumes corresponds to vectors for which at least two coordinates are positive. (b) Image labeled with the color-name descriptor. . . . . 48

3.3 Color-shade descriptor. An image from the MIT dataset (a), its color distribution (b), and the ridges detected by the RAD method (c). . . . . 49

3.4 Block diagram of our method for intrinsic image estimation. . . . . 50

3.5 Schema of the three different scenarios we considered in the description of the pairwise potential. . . . . 52

3.6 Schema of our Markov random field. (a) Each pixel  $x_i$  has two observations, one from the color-name descriptor ( $ND$ ) and another from the color-shade descriptor ( $SD$ ). The continuity between two pixels is enforced in the pair-wise potential. The width of the lines represents the weight we assign to each of the edges in order to penalize label discontinuities. These weights are locally modified according to the different cases defined by the color-shade observations: (b) the pixels belong to different ridges (Case A). We assign a small weight to this edge, since no continuity is expected between these two pixels in the reflectance image. (c) The pixels belong to the same ridge but the line they form is not parallel to the ridge (Case B). We assign a medium weight to this edge. We want to penalize different neighboring labels moderately as a consequence of the unclear conclusion we draw from the color-shade descriptor, allowing certain flexibility for label changes in this scenario. (d) The pixels belong to the same ridge and the line they form is parallel to the ridge (Case C). We assign a big weight to this edge, since we do not want these pixels to have different labels in the reflectance estimate. Notice that the weights always verify  $\alpha \leq \beta \leq \gamma$  and  $\alpha \ll \gamma$ . . . . . 53

3.7 (a) Original Image. Different reflectance and shading estimates are shown: (b) Only local information of the color-name descriptor is used. (c) Output of the MRF where semi-local information of the ridge observations has been combined with local information of the color-name descriptor. (d) Adding global scene coherence to the output of the MRF. 55

4.1 Illustration of the multiple variables that model the spectral radiance. 68

4.2 Overview of the proposed model. Modeling the effects of the illumination and the sensors in the acquired images allows intrinsic image methods to obtain absolute reflectance images  $I_{Ref}^a$ . . . . . 73

- 4.3 Mean angular error for different commercial cameras in multiple scenarios. The most information we have about the scene, the best our reflectance estimate will be. Scenario 4 always assumes canonical illumination and standard sensors, like most existing intrinsic image models. Scenarios 1 and 2 represent our model when there is some knowledge about scene illumination and camera sensors. . . . . 79
- 4.4 Image categories: From left to right, the Golden Gate bridge (San Francisco), the Statue of Liberty (New York City), the Kinkakuji temple (Kyoto), the church of Sagrada Família (Barcelona), the Uluru Rock (Australian desert) and St. Basil’s cathedral (Moscow). . . . . 80
- 4.5 Results on 3 pairs of natural images. At each row, in the first column we see a pair of original images. In the second column the Grey-Edge algorithm [149] has been applied. In the last column, we see the images of the second column after their camera sensor effects have been removed. The chromatic angular errors are expressed in degrees. . . . 82
- 4.6 Mean angular error among pairs of images from our set of natural images corresponding to the scenarios 2, 3 and 4 previously defined. The Shades of Grey algorithm [52] has been used to estimate the illuminant, but similar results are obtained when using other color constancy methods, such as Grey-Edge [149]. . . . . 83
- 4.7 Example of the influence of the color of the illuminant and the camera sensors on intrinsic images. (a) and (e) are images of a given landmark, taken with different cameras under different illumination conditions. (b) and (f) are the images in (a) and (e) respectively, after removing the effects of the illuminant and the camera sensors. (c),(d),(g) and (h) are the estimated intrinsic images of (a),(b),(e) and (f), respectively. Chromatic angular error values are given in degrees. . . . . 84
- 4.8 Results showing the average chromatic difference for a synthetic object rendered under different lighting and sensor conditions. The first and second rows share the same camera sensors but not lighting conditions. The second and third rows share the scene illuminant but not the camera sensors. Ideally all 3 images should have the same intrinsic reflectance image. Numbers under each column are the average difference between images as measured by angular differences in sRGB space. (a) Original Image. (b) Reflectance estimates of the method of Serra *et al.* [133]. (c) Reflectance estimates of the method of Barron *et al.* [19]. (d) and (e) Removing the effects of the color of the illuminant and the camera sensors from input images minimizes the differences between the three reflectance estimates for [133] and [19], respectively. 85

4.9	Results showing the average chromatic difference for two objects acquired in the laboratory under different lighting conditions. Ideally, for each object, all 3 images should have the same intrinsic reflectance image. Numbers under each column are the average chromatic difference between images as measured by angular differences in sRGB space. (a) Original Image. (b) Reflectance estimates of the method of Serra <i>et al.</i> [133]. (c) Reflectance estimates of the method of Barron <i>et al.</i> [19]. (d) and (e) Reflectance estimates for [133] and [19], respectively, after removing the effects of the color of the illuminant and the camera sensors. . . . .	86
4.10	Variability of relative and absolute reflectance of an object under two different illuminants. (a) Original image. (b) Theoretical relative reflectance. (c) Theoretical absolute reflectance. . . . .	87
5.1	Objects in our Dataset. . . . .	90
5.2	Schema of our laboratory. We illustrate the position of the two cameras, the three light bulbs and the object (in the center of the room). The color filters that we used to change the color of the illuminant are also represented, as colored regions, close to the light bulbs. . . . .	92
5.3	Example of the ground truth data we can provide for a single object and camera. . . . .	93
5.4	Removing the effect of the illuminant and the camera sensors increases the similarity between the different estimated images. In the upper row we see an object from the dataset acquired with the Nikon D5200 camera and under a reddish illuminant, while in the lower row we see the same object acquired with the Sigma Foveon D10 camera under a bluish illuminant. In the first column we see the original input images. In the second column we observe the reflectance estimates according to Gehler's method [60] when we assume white light and sRGB sensors. Finally, in the third column we see the reflectance estimates of the same method when information about the camera model and the illuminant of the scene is available and the effects of both factors have been removed. . . . .	96
5.5	Comparing different rendering methods: <i>direct lighting</i> (a) and <i>photon mapping</i> (b). . . . .	99
5.6	Single object scenes included in our dataset. . . . .	101
5.7	Complex scenes included in our dataset. . . . .	102
5.8	Two examples of ground-truth decomposition. For both rows, from left to right: the rendered scene, reflectance component, and shading-illumination. . . . .	102
5.9	Two examples from the dataset under different illumination conditions. For both columns, from top to bottom: white illuminant, single-colored light, and two distinct colored illuminants. . . . .	103
B.1	Spectral power distribution of the 10 illuminants used in the experiment with synthetic data. (a) Planckian illuminants. (b) Non-Planckian illuminants. . . . .	116

C.1 Images of single objects in our synthetic dataset. . . . . 126  
C.2 Images of complex scenes in our synthetic dataset. . . . . 127

# Chapter 1

## Introduction

Computer vision is the scientific field that focuses on computationally emulating the processes that happen in the human visual system by acquiring, processing, analyzing and understanding images. These images usually contain multiple complex effects, such as cast shadows, specularities or interreflections, which are caused by the geometry of the scene, the position of the light sources and the material of the objects. Moreover, the color of the illuminants and the response of the camera sensors also affect the image values resulting from the acquisition process [111]. Isolating and extracting these effects from images would benefit most computer vision tasks, such as image segmentation, object recognition, video tracking, scene reconstruction, etc.

For instance, common object recognition problems, such as pedestrian detection in autonomous vehicles or license plate detection in traffic monitoring are strongly affected by brusque changes in illumination [8, 83]. Being able to remove these illumination effects (cast shadows, specularities, color of the illuminant, etc.) would dramatically simplify the problem and increase the accuracy of most of the existing approaches in the field [114]. Moreover, isolating the information about the illumination would be extremely useful in order to analyze different physical properties of the scene, such as the number, color and direction of light sources, or the geometry, the materials and relative position of the objects.

Images describing a single physical feature of the scene, such as the effects described above, are called *intrinsic images* [23]. In this thesis we will analyze the problem of intrinsic image estimation from different perspectives, including the theoretical formulation of the problem, the visual cues that can be used to estimate the intrinsic components and the evaluation mechanisms of the problem. To this end, we first need to introduce some background.

In this chapter, we will briefly discuss how three-dimensional scenes have been represented in planar surfaces throughout history. Then, we will briefly mention how the human visual system works. We will sketch the perceptual processes that are supposed to occur in our brains. We will review some of the layer decomposition models which aim at explaining the perceptual processes of our brain and which serve as a theoretical background for computer vision techniques. Finally, we will summarize the seminal work of Barrow and Tenenbaum on intrinsic images [23],

which is at the basis of this thesis.

## 1.1 Visual Descriptions of Scenes

An image, or picture, is a two-dimensional visual representation of a three-dimensional scene. When the scene lies in a three-dimensional space, the image is a projection of the scene on a plane or on a curved surface. These graphical descriptions, either in the shape of drawings, paintings or photographs, have played an important role during all the history, immortalizing both quotidian actions and special events. The evolution of these arts deserves some attention, since it has usually been parallel to our understanding of the human visual system [69].

The visual description of scenes on planar surfaces has been recurrent ever since humans inhabited caves. Evidence of this are the multiple cave paintings dating back to paleolithic and neolithic periods [127]. These paintings used very rudimentary techniques that depicted animal silhouettes in hunting scenes. Drawing and painting have constantly evolved, and the development and enhancement of techniques to represent visual cues such as depth, volume or transparency have progressively added a strong sense of realism to these pictures [67].

Until the end of the Middle Ages, basic techniques such as occlusion, scale or the relative position in the plane were used to describe the third dimension (*i.e.* volume and depth) [132, 67]. More complex techniques capable of conveying a sense of transparency were also developed in order to represent materials such as silk [130]. Examples of these primary techniques can be seen in Figure 1.1.

During the Renaissance new techniques were explored, achieving a more realistic sensation of volume and depth in paintings [72, 67]. Linear perspective, where the parallel lines of an object converge into a given vanishing point and the sizes of the figures are reduced according to distance, constituted a significant step forward to represent depth. In order to enhance the sensation of volume in objects, emphasis was also put on illumination (*i.e.* shading and cast shadows), textures and transparency. Examples of this are painting techniques such as *sfumato*, which produces soft shadows and imperceptible transitions between colors, or *chiaroscuro*, which creates strong contrasts between light and dark. Examples of these techniques are shown in Figure 1.2.

Until the 19th century, drawing and painting were mainly intended to faithfully represent real scenes, and were therefore concerned with providing a strong sense of realism. The development of photography, however, had a big impact in these arts and quite soon different abstract art movements, which focused on representing subjective feelings rather than physical objects, appeared.

The first camera device was built at the beginning of the 19th century, while color photography appeared in the 20th century, and it was not until the end of this same century that modern digital photography came out. Although the goal of many photographers has been to express their subjective perceptions and emotions [125], we are interested in the ability of camera devices to represent objective features, which can be measured physically in a scene.

Although many issues influence the final pixel values we observe at each point in



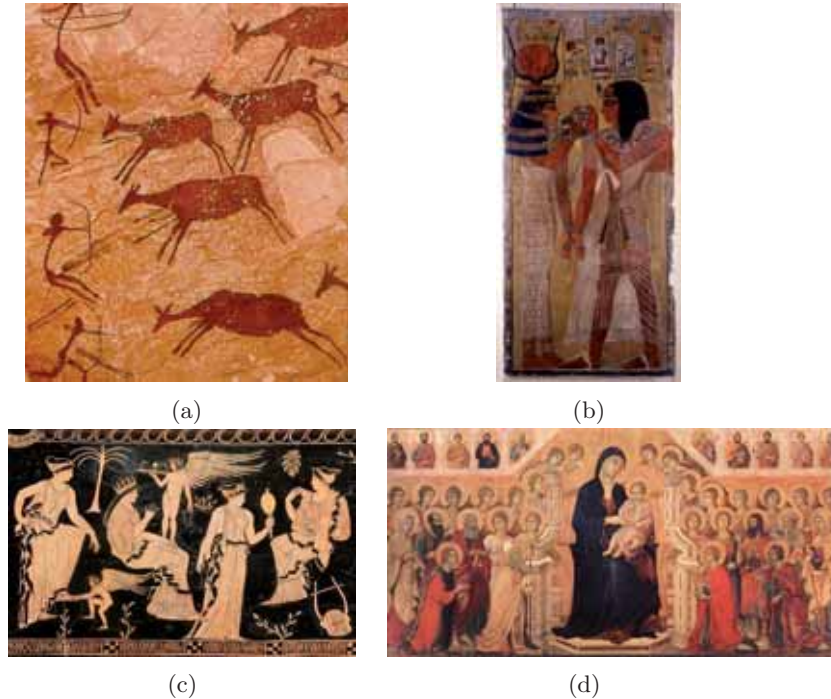


Figure 1.1: Examples of techniques used to represent depth before the Renaissance period. (a) A cave painting found in Valltorta, Valencia (Spain), where relative position has been used to represent depth. (b) Transparent materials were represented in ancient Egypt in the painting *The goddess Hathor welcomes Sethos I*. Occlusion was also used in this painting to represent depth. (c) An example of ancient Greek art where occlusion and relative position are used to convey a sense of depth. (d) A sample of Byzantine art, the altarpiece *Maestà of Duccio*, where occlusion, scale and relative position are used.

the image, some factors are of central interest: the position and nature of the light sources in the scene, the reflectance properties of the objects and the responses of camera sensors [111]. The direction of the illuminant and the geometry of the objects in the scene result in cast shadows and shading variations, while the color and power of the illuminant affect the colors we perceive in the images. The materials of the objects are also fundamental during the image acquisition process, since different materials have different reflection properties. For instance plastic and metallic materials will result in specularities in the image while wooden materials are diffuse. Finally, if we take a picture of a single scene using multiple camera devices, the corresponding images will have different color values. This is due to the response of the sensitivity functions of the camera sensors, which are different for each camera model. The acquisition process and the importance of these three factors will be further detailed in Chapter 4.

After being acquired, digital images go through a quantization process in order



Figure 1.2: Examples of techniques that came out during the Renaissance period. Leonardo da Vinci's painting *L'Ultima Cena* (a) is a clear example of linear perspective. Sfumato was used in one of the most famous art pieces, Leonardo da Vinci's painting *La Gioconda* (b). Giovanni Baglione's painting *Amor sacro e amor profano* (c) provides a great example of chiaroscuro. These techniques convey a vivid sensation of shape and volume.

to be encoded as standard digital elements which can be read and viewed in any computer device. This quantization process performs a discretization of the image values both geometrically and colorimetrically, depending on the desired resolution (*i.e.* size of the image) and color description of the image [68]. The acquisition and quantization processes are both schematically illustrated in Figure 1.3.

So far we have seen that three-dimensional scenes have been represented in planar surfaces throughout history. In drawing, painting and photography, information of the scene originating from multiple factors (*i.e.* geometry, illumination, reflectance of the objects, etc.) is combined in a single image.

As mentioned above, computer vision aims at emulating the human ability to recognize objects and actions in images. However, images values are the result of visual information coming from multiple sources. Isolating the information coming from each of these sources would benefit most computer vision tasks. In this work, given an image, we want to recover information from the multiple factors that produced it. Although it may seem a hard and ill-posed problem, it is important to observe that

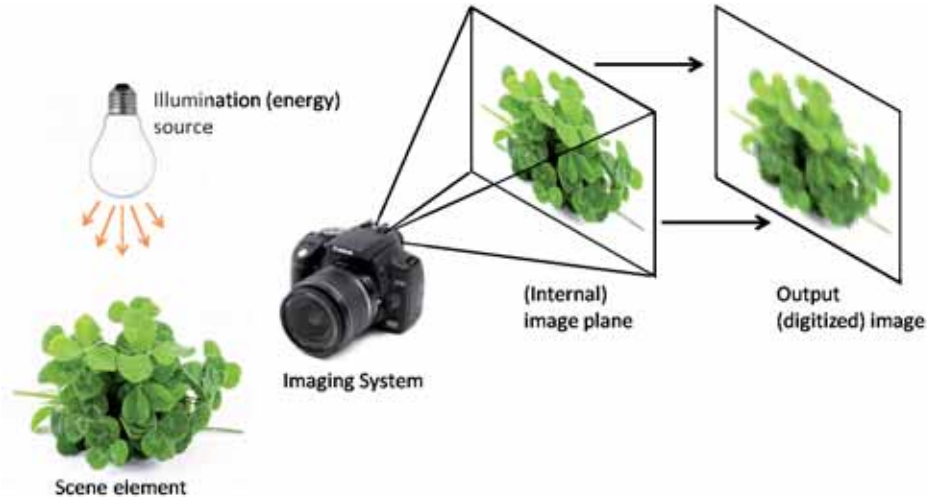


Figure 1.3: Image acquisition process. The light source, the materials of the objects and the response of the camera sensors influence the resulting image values. Afterwards, during the quantization process, geometric and colorimetric discretizations are applied to the image.

whenever we refer to drawings, paintings or photos, human beings develop the ability to interpret these images and recognize and isolate specific physical properties of the scene. For example, we can estimate the reflectance of the objects, even when their appearance is affected by an illuminant which modifies their colors. We are also able to interpret the shape of an object or the depth in a scene, even when we just see a simple drawing or sketch.

In computer vision, different visual cues are used to recognize these specific physical properties. It is assumed that small intensity variations are usually caused by shading, while sharp luminance differences mostly result from reflectance changes. Nonetheless, in order to know which visual cues are the most useful for this purpose, we need to understand how our visual system works.

## 1.2 Outline of the Human Visual System

The human visual system is responsible for detecting and interpreting information from visible light and its interaction with objects in order to build a representation of the surrounding environment [147, 92, 115]. The retina in the human eye samples the world in a process similar to the process of image acquisition in photography. On the other hand, the visual cortex in our brain is responsible for interpreting the signal transmitted by the retina.

Our retina has two different types of photoreceptor cells: rods and cones. Cones can be subdivided into three types, namely S-cones, M-cones, and L-cones, where the capital letters refer to their peak sensitivities in the short, medium, and long

wavelength regions of the visible spectrum (*i.e.* 380 - 750 nm). The output of each type of cone provides measurements of incoming light intensity over a broad range of wavelength, but with peak sensitivities at different wavelengths. Having three cone types with broadly tuned and overlapping wavelength sensitivities provides measurements of the luminance spectrum at each location in the retinal image. Rods, on the other hand, are much more sensitive to light than cones, and are mainly responsible for vision in dim light. The sensitivity responses for the different photoreceptor cells are shown in Figure 1.4. In photography, the camera sensors play the role of the photoreceptor cells that can be found in the human retina [147, 92, 115].

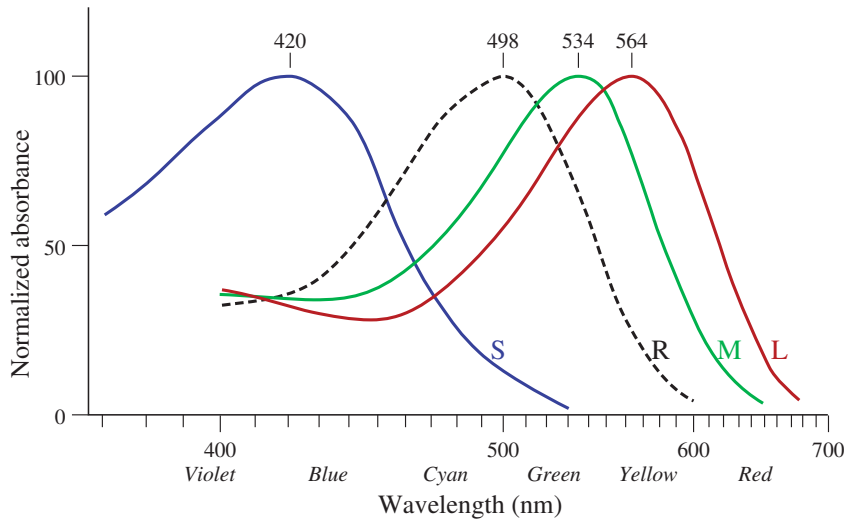


Figure 1.4: Response of the different photoreceptor cells in the retina of human beings with normal color vision. R indicates rods. S, M and L indicate the different types of cones.

Still in the retina, visual information from the photoreceptors is collected by the organization of center-surround receptive fields of the ganglion cells, which provides a way to detect contrast. The optic nerve connects the retina with the visual cortex, acting as a continuous mapping of retinal points to corresponding cortical points. In the brain, different areas of the visual cortex are specialized for processing particular types of information such as movement, depth, etc. [147, 92, 115].

Although the way that cells work in the retina is well-understood, the analysis which takes place in the visual cortex is not fully understood [155, 115]. Psychologists have been trying to unveil the perceptual processes happening in the human brain, studying visual effects such as brightness induction [129] (*i.e.* a phenomenon by which the perceived luminance of an area is modulated by the luminance of the surrounding areas), lightness constancy [113] (*i.e.* the ability to perceive the same brightness in objects under different conditions of illumination) and color constancy [46, 54] (*i.e.* the ability to have a stable perception of the colors of objects, irrespective of the color of the illuminant). Examples of these effects can be seen in Figure 1.5.

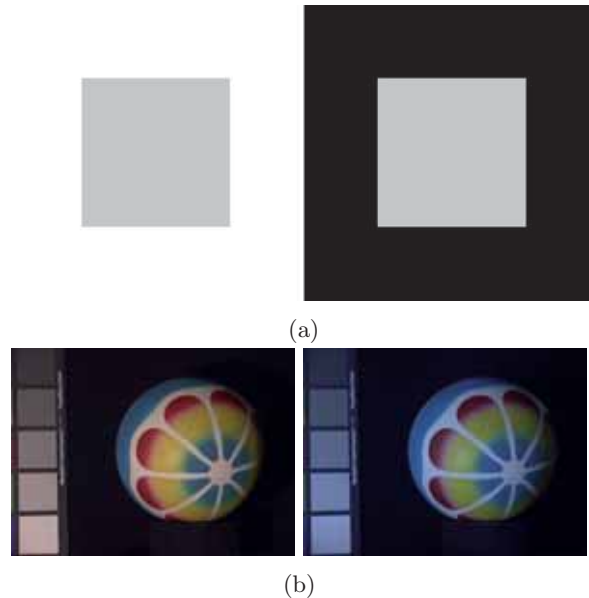


Figure 1.5: Examples of brightness induction (a) and color constancy (b). In (a), although both grey squares have exactly the same color, the square surrounded by black is perceived brighter than the square surrounded by white, a phenomenon referred to as brightness induction. In (b), we see the same scene under two different illumination conditions: A reddish light in the image on the left and a bluish light in the image on the right (data from [17]). Although these illumination conditions result in different pixel values, we perceive the same colors in both images. This phenomenon is referred to as color constancy.

The layer decomposition of images into a set of intrinsic physical components is one of the models that psychologists have used to describe the way our visual system perceives and interprets visual information. The problem of intrinsic image estimation, which is the basis of this thesis, focuses on computationally emulating this perceptual model. We provide a brief background on this perceptual model in the next section.

### 1.3 Background on Human Visual Perception

We have already mentioned that once in the brain, information is processed in different specialized areas of the visual cortex [155, 115]. Psychologists have been trying to unveil these processes.

In psychology, perception is defined as the neurophysiological processes by which an organism becomes aware of and interprets external stimuli [84]. In this section we focus on the role that intrinsic images play in human visual perception. Our aim in this thesis is to study how these intrinsic image models can be computationally implemented to emulate human visual perception.

The starting point for the study of intrinsic image models can be considered the works on simultaneous color contrast of Helmholtz [75] and Hering [76] in the second half of the 19th century. In their respective works, these authors drew attention to the importance of layer perception, which has influenced many posterior works. Layer perception refers to the idea that the visual system decomposes images into multiple layers (or intrinsic images) in order to distinguish light from material [95].

Following Helmholtz and Hering ideas about color contrast, some authors assumed that human perception is governed by edge information, and, therefore, that there must be some process whereby reflectance changes are distinguished from illumination variations. Sound contributions in this direction are the works of Land and McCann [104, 103], and Gilchrist and colleagues [64, 65, 66]. These approaches suggested that this separation process extracts luminance variations from the input image, classifies these luminance derivatives as either being caused by reflectance or illumination changes, and finally integrates these luminance edges according to their classification, obtaining a reflectance image and an illumination image.

Land and McCann, in their well-known Retinex theory [104, 103], defined a two-dimensional Mondrian world (*i.e.* a world made of color patches) where illumination gradients cause smooth intensity variations and reflectance variations produce step intensity changes. The Retinex theory [104] has been the basis of many posterior works in intrinsic image decomposition, and it will be further discussed in Chapter 2.

Gilchrist *et al.* observed in [64, 65] that in three-dimensional scenes, sharp intensity changes can arise from either reflectance or illumination variations. Though the authors do not specify an edge-classification technique, they observed that edge intersections may play an important role in such classification. Intersections where an illumination edge crosses a reflectance edge satisfy a ratio-invariance property, which is illustrated in Figure 1.6.

Metelli, also influenced by Helmholtz and Hering's ideas, defined the theory of color scission in his classical work about the perception of transparency [120]. Color scission explains transparency as a case of perceptual color-splitting or layer decomposition. Metelli also pointed out that the main cues for the perception of transparency are to be found in figural and chromatic conditions (see Figure 1.7).

All these authors agreed on the importance of layer decomposition in human perception, which inspired Barrow and Tenenbaum in their definition of intrinsic images [23]. Their seminal work is summarized in the following section (Section 1.4).

Psychologists have broadly studied the human ability to perceive and isolate illumination effects from reflectance, an aptitude called lightness or color constancy. In their perceptual studies on intrinsic image models, there are three perceptual 'layers' (*i.e.* intrinsic images) which are thought to be critical to vision: reflectance, illumination and transparency [9, 96].

In order to emulate computationally how human visual perception works, we need to understand some of the ideas in the works on perception of intrinsic images. Kingdom, in [96], observes that these perceptual models on intrinsic images are, in fact, compilations of demonstrations showing the influence of illumination and transparency on surface lightness and brightness. The author proposes two important factors to classify these works.

The first factor is related to the cues used by human vision to identify the presence

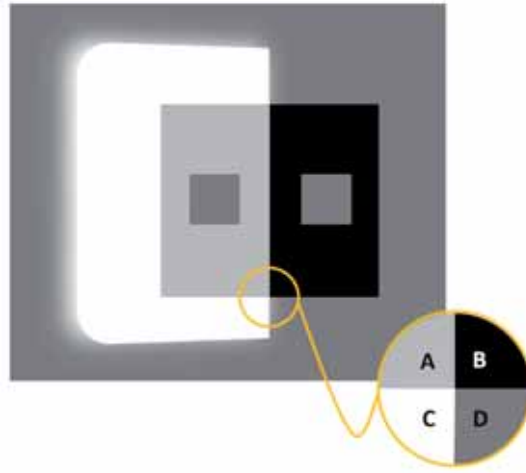


Figure 1.6: Representation of the ratio-invariance property at edge intersections described in [64]. An illumination edge (the vertical one) dividing regions A and C from regions B and D crosses a reflectance edge (the horizontal one) separating regions A and B from regions C and D) and the ratio-invariance property,  $\frac{A}{B} = \frac{C}{D}$ , holds.

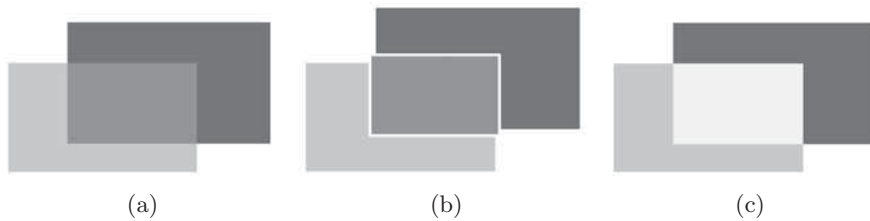


Figure 1.7: The perception of transparency [120]. The transparency perceived in (a) disappears when the figural unity is broken (b) or when certain color conditions are not met (c).

of non-uniform illumination and transparency. Among these cues, we find luminance relations such as the ratio-invariance at intersecting edges described by Gilchrist in [66], gradient classification, popularized by Land and McCann in [104], or contrast in luminance edges [110, 47]. We also find figural relations such as junctions, used in [66, 9], and straightness properties [110]. Other common cues such as depth, color, texture and motion are also described (see Kingdom's review [95]).

The second factor is concerned with the effects of perceived illumination and transparency on lightness and brightness. Different works have shown how depictions of transparency [9, 110], shading and shadows [10, 11, 109], and figure-ground relationships [12] can profoundly influence brightness perception. Kingdom concludes that existing perceptual intrinsic image models do not explain how layer decomposition and luminance values are combined to compute lightness. Nonetheless, these models have explored the visual cues involved in layer decomposition and the role that

transparency and non-uniform illumination play on our perception of illumination (brightness) and reflectance (lightness).

Some attention must be given to the works of Adelson [9, 11]. His works on intrinsic images are especially interesting in the field of computer vision, since he is one of the few authors who has linked both perceptual and computational intrinsic images models.

In [9], Adelson presented illusions that show how three-dimensionality and transparency interpretations can affect brightness judgements. These illusions cannot be explained with low-level mechanisms such as the mechanisms involved in the Retinex (see Figure 1.8). The author concluded that any model predicting brightness phenomena should use sophisticated mechanisms that decompose the image into a set of intrinsic images representing reflectance, illumination and transparency. Adelson's observations on the Retinex theory are significant since the Retinex has been the basis for many works on intrinsic image decomposition in the field of computer vision. Moreover, while most computational methods for intrinsic image decomposition have focused on decomposing input images into their reflectance and illumination components, transparency has been thoroughly ignored.

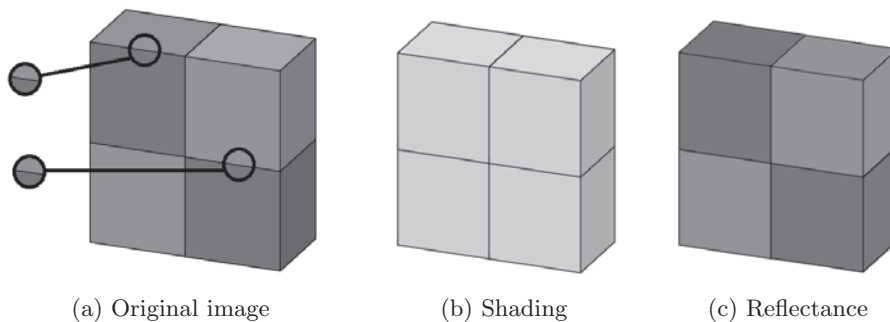


Figure 1.8: The influence of 3D interpretation on brightness and lightness perception (image adapted from [10]). The edges in (a) would be equally classified by local methods such as the Retinex [104], although the upper one is a result of a variation in illumination (b), while the lower one is due to a change in reflectance (c).

Adelson and Pentland [11] observed that any mechanism for achieving an intrinsic image decomposition must make assumptions about regularities in the scenes. The authors described an algorithm that could interpret simple polyhedral images using a hierarchical model based on cost functions that enhance the most likely explanation of the image according to a set of predefined independent rules on shape, lighting and reflectance. Their idea is illustrated in Figure 1.9, where different combinations of shape, lighting, and reflectance result in the same image. However, according to a set of predefined independent rules about the scenes (*i.e.* prior knowledge), one explanation of the intrinsic composition of the image is more probable than the others. For instance, in this example we see that one possible explanation of the resulting image is given by a flat object illuminated from the viewer's direction where all luminance variations occur in the reflectance component (first row). However, the other explanations (second and third rows), where some of these luminance



are caused by the geometry of the scene (*i.e.* shading changes), seem far more intuitive. The perceptual model of Adelson and Pentland [11] was presented as a computational method for intrinsic image decomposition in [141].

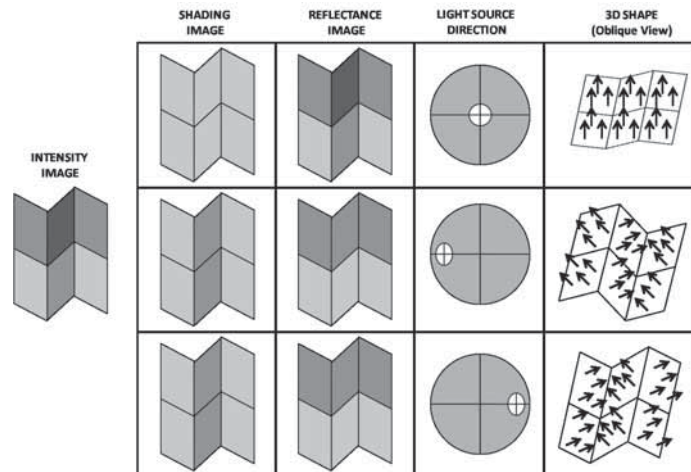


Figure 1.9: Image adapted from [11]. Different illumination, shape and reflectance conditions result in the same image. Are some of these combinations more likely than others?

## 1.4 Intrinsic Images in Computer Vision

The problem of intrinsic image decomposition in computer vision was first defined by Barrow and Tenenbaum in [23]. The authors introduced the recovery process of “intrinsic images” from one or multiple intensity images (although color images are more commonly used nowadays). Intrinsic images were defined as images containing a single intrinsic characteristic of a scene. The primary intrinsic characteristics that the authors proposed to be recovered included material reflectance, distance or surface orientation, and incident illumination (*i.e.* shading). They also named others such as transparency, specularity or luminosity.

The main problem in recovering intrinsic scene characteristics from an intensity image is that the information in the image is confounded, since each intensity value combines all the intrinsic characteristics of the corresponding scene point. Although the information in the input image may seem insufficient, it is undeniable that humans possess the ability to estimate physical characteristics of a scene throughout a wide range of viewing conditions, even if the scene is unfamiliar.

Barrow and Tenenbaum examined the computational nature of the recovery process to determine whether such a design is really feasible. The authors first defined a simple world or experimental domain, making simplifying assumptions about the scene, the illumination, the viewpoint, the sensors, and the image-encoding process. In this simplified domain, images consisted of regions of smoothly varying intensity

bounded by step discontinuities. The nature of these edges was studied and a catalog for edge classification was proposed.

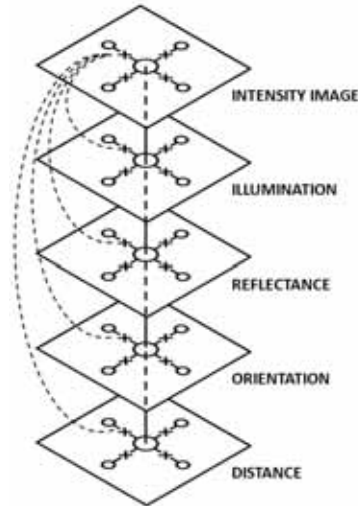


Figure 1.10: Computational model for the recovery of intrinsic image presented in the work of Barrow and Tenenbaum [23]. In this model, an input intensity image (first layer) is decomposed into multiple intrinsic images (subsequent layers) using intra-image continuity (circles), inter-image constraints (vertical lines) and further local processes (X marks) in each layer to inhibit continuity constraints.

Figure 1.10 shows a schematic of the computational model for the recovery of intrinsic images defined by Barrow and Tenenbaum in [23]. Given an input intensity image, represented in the first layer of the model (Figure 1.10, top), intensity edges are identified and interpreted according to the proposed catalog for edge classification. According to these interpretations, edges are defined (dashed curved lines) in the multiple intrinsic images.

A first set of processes (circles) modifies the values locally in each layer in order to enforce intra-image continuity and limit constraints. Then, the consistency of the multiple intrinsic image values is enforced in a second set of processes (vertical lines) by using inter-image photometric constraints. Finally, a third set of processes (X marks) is applied to each layer to insert and delete edge elements in order to locally inhibit continuity constraints. These processes interact continuously to improve the initial edge interpretation until accurate intrinsic scene characteristics are recovered.

For instance, imagine that we want to decompose a color image into its shading and reflectance intrinsic components. We can use the Retinex assumption which states that small luminance edges are the result of shading variations, while sharp luminance edges are usually caused by reflectance changes. Thus, in that case, we define a threshold and use it to classify the luminance derivatives of the input image (this classification corresponds to the curved dashed lines in the model). Once we have classified these edges as being caused by either shading or reflectance variations,

we use assumptions about the regularities of the scenes, such as the smooth variation of shading, to locally modify the values of our intrinsic images and enforce intra-image continuity (circles) in the shading component. Now, given that we decompose the image into two different intrinsic images, let us assume that the product of these images must be equal to the input image. Using this information we know that when we modify one intrinsic component we must modify the other accordingly in order to enforce inter-image photometric constraints (vertical lines). Finally, we also use information about the chromatic derivatives of the input image in order to locally inhibit (X-marks) continuity constraints that were enforced by our previous shading smoothness assumption. Since intensity variations are canceled in the chromatic image, we can now decide whether some of the edges that we classified as being caused by reflectance changes are, in fact, sharp shading variations caused by a cast shadow. This information modifies the nature of the edges that we classified at the beginning.

Barrow and Tenenbaum discussed how this model can be extended to work in more complex domains and also studied the consequences of relaxing the different assumptions of their initial oversimplified world. They observed that phenomena such as specularities and transparencies, present in many real scenes, could be helpful in the recovery process for completely describing the scene since they represent additional intrinsic characteristics. It was also mentioned that the described framework could be extended to accommodate additional sources of information about the scene such as input color images.

The work of Barrow and Tenenbaum [23] was strongly influenced by the previous works of Horn [80, 81, 82] and Marr [117]. Horn studied the physical basis of image intensity variations [82], and designed techniques for determining surface lightness [80] and shape from shading [81] in simplified domains. Barrow and Tenenbaum [23] studied the constraints and assumptions underlying Horn's recovery techniques with the idea that they could be integrated in a more general method which could be used in more complex scenes. Marr described a layered organization for a general vision system in [117]. In his work, the first layer consists of a symbolic representation of the edges and shading in an intensity image, which is used at the next layer to derive the three-dimensional structure of the image. This structure is analogous to Barrow and Tenenbaum orientation and distance intrinsic images. However, Marr focused on understanding the nature of individual cues, perfecting the symbolic representation of the edges before undertaking any higher level processing, while Barrow and Tenenbaum focused on understanding the integration of multiple cues, attempting to immediately assign three-dimensional interpretations to intensity edges.

Following the work of Barrow and Tenenbaum, different authors focused on decomposing images into their shading and reflectance components by making simplifying assumptions of the world and classifying image derivatives (we review these methods in Chapter 2). Recovering these two characteristics from a single image  $I$  amounts to estimate a reflectance image and a shading image such that

$$I(x, y) = I_{Shad}(x, y) \cdot I_{Refl}(x, y), \quad (1.1)$$

where  $I_{Shad}$  represents the amount of reflection arriving to the  $(x, y)$  point of the image from a specific point of the object surface and  $I_{Refl}$  describes how the light is reflected by the corresponding point of the object. This problem is clearly ill-posed

since it is underconstrained: the unknowns outnumber the equations.

One of the drawbacks of this formulation is that it does not model the effect of either the illuminant of the scene or the camera sensors. As we have previously mentioned, both factors influence the pixel values of the images after the acquisition process. As a result, these factors strongly affect the reflectance estimates of any method for intrinsic image decomposition. However, they are usually ignored. Here we briefly introduce the related topics of color constancy and camera sensor calibration in Subsections 1.4.1 and 1.4.2 respectively. In Chapter 4 we propose a new generalized framework for intrinsic image decomposition where the effects of the illuminant of the scene and the camera sensors have been included in the formulation of the problem. We also show how color constancy and sensor calibration models can be used to improve the estimates of existing intrinsic image decomposition approaches.

In a completely different direction, another problem that early authors in intrinsic image decomposition faced was the lack of standard evaluation mechanisms which prevented them to quantitatively compare their methods against other approaches. In Subsection 1.4.3 we briefly introduce the problem of intrinsic image evaluation. In Chapter 2 we review the current evaluation mechanisms and, in Chapter 5, we present two new datasets for intrinsic image evaluation.

### 1.4.1 Color Constancy

Human beings have the ability to isolate and correct the effects of the color of the illuminant [54] on the different reflectances in the scene. This ability is called color constancy, and the way it works is still unclear, even if several neural mechanisms contributing to color constancy have been singled out [101]. Figure 1.5(b) provides a simple example of color constancy.

Observe that acquiring two images of a single scene (with the same camera device) using two different illuminants results in two images with different color values. In the field of computer vision, filtering out the effects of the light source from images is fundamental for any color-based technique. Several computational color constancy approaches exist [52, 149, 63, 79]. The goal of these methods is to produce a corrected image where the effects of the color of the illuminant have been removed, or in other words, estimate how the image would look under a canonical “white” light (*i.e.* illuminant representing midday sunlight). In order to achieve this objective, many approaches estimate the color of the illuminant and then perform a color transformation on the input image, resulting in a canonical image. Notice that in order to perform such a color correction on the whole image, these methods assume that there is a single light source and its color is distributed homogeneously across the scene. However, this assumption does not hold when there are multiple light sources in the scene or interreflections are found in the images.

Finally, research in the field of color constancy has also devoted attention to the effects of the imaging sensors for the final recovery of the canonical illuminants [152].

As mentioned above, in Chapter 4 we further discuss the influence of the illuminant in the problem of intrinsic image decomposition and show how color constancy can be used to achieve a better intrinsic reflectance estimate which is invariant to the color of the illuminant.

### 1.4.2 Sensor Calibration

Camera sensors are responsible for capturing light and converting it into pixel values. This process is achieved by color-separation mechanisms, which depend on the response of the sensitivity functions of the sensors. For this reason, if we simultaneously took a picture of the same scene with two different cameras from the same point of view and using the same settings, the pixel values of the resulting images would not be the same.

As it happened with the color of the illuminant, camera sensors also affect the pixel values that we observe in images. Two images acquired with different camera models under the same illumination conditions will have different color values. Therefore, filtering out the effects of these camera sensors from images is also fundamental for any color-based technique in computer vision. There exist several computational models of camera calibration that estimate the sensitivity functions of the camera [16, 124]. Once these sensitivity functions have been estimated, it is possible to perform a color transformation on the input image in order to obtain a corrected image where the effects of the camera sensors have been removed, or in other words, estimate how the image would look under a set of standard sensors. These sensor transformations are often modeled with 3-by-3 matrices [85].

In Chapter 4 we further discuss the influence of the camera sensors in the problem of intrinsic image decomposition. We show how any knowledge about the camera model used to capture the image can be used to achieve a better intrinsic reflectance estimate which is invariant to the camera sensors.

### 1.4.3 Intrinsic Image Evaluation

As mentioned above, one of the difficulties that the first computational methods on intrinsic image estimation faced was the lack of standard ground truth datasets and metrics for intrinsic image evaluation. In fact, most early methods only showed a few qualitative examples.

The appearance of the MIT dataset [71] allowed many posterior methods to be quantitatively evaluated and compared against other methods. However, building such a dataset proved to be challenging, and the small number and variety of ground truth data it contains have forced some recent methods to evaluate their results with other datasets which have not been specifically built for the purpose of intrinsic image evaluation [38, 159].

Given the actual complexity of building datasets for intrinsic image evaluation using natural scenes, different alternatives such as crowdsourced datasets [28] and synthetic image datasets [24], have been proposed. In Chapter 2 we present a thorough review of the different datasets and metrics used so far in the field. In Chapter 5 we present two new datasets for intrinsic image evaluation.

## 1.5 Scope of the thesis

The main goal of this thesis is to analyze the influence of color features in the problem of intrinsic image estimation and to propose multiple improvements in different areas

of the problem. We describe a new method for intrinsic image estimation, define a new theoretical model which generalizes previous formulations of the problem and release two different ground truth collections for intrinsic image evaluation.

In Chapter 2 we first provide a thorough review of the existing methods for intrinsic image decomposition, as well as the different datasets and metrics that have been used so far in the field. We analyze how the simplifying assumptions about the scenes have modified the formulation of the problem and also how different information cues based on regularities about the scenes have been used to estimate intrinsic images. We finally discuss the evolution of the problem and anticipate future developments in the field.

Although color information has been frequently ignored in computer vision [62], it has proven to be extremely useful in the decomposition of intrinsic images. In Chapter 3 we present a computational method for intrinsic image decomposition which, given a single input image, estimates its intrinsic reflectance and shading components using information from different color cues. The first cue is based on the semantic description of color used by humans (*i.e.* based on color names) and provides a more robust description of the reflectances than standard color spaces. The second cue is based on an analysis of color distributions in the histogram space and provides a consistent description of surfaces sharing the same reflectance, overcoming local problems that the color name descriptor may find in shadowed or near highlight regions of the image. A probabilistic framework is used to combine information from both color descriptors.

In most methods for intrinsic image decomposition, including ours, authors have assumed white light in the scenes and have completely ignored the effect of camera sensors in images. However, as explained above, both the illuminant of the scene and the camera sensors strongly influence the resulting pixel values during the image acquisition process. In Chapter 4 we analyze the theoretical formulation underlying the decomposition problem and propose a generalized framework where we model the effects of both camera sensors and illuminant color. Our framework extends previous formulations. Moreover, we show how removing these photometric effects from input images improves the results of different methods for intrinsic image decomposition.

In order to validate our framework we have built a calibrated dataset which includes ground truth information about the illuminant of the scene and the camera sensors. This new dataset is presented in Chapter 4. The evaluation of intrinsic images has continuously evolved during the last decade. Although different datasets have been released, building these datasets has proved to be a challenging task. Current ground truth collections present some drawbacks, such as the small number and diversity of scenes or the lack of ground truth information for specific intrinsic components (*i.e.* the depth or surface orientation of the objects in the image, the color and direction of the illuminant, etc.). In Chapter 5 we also present another new dataset for intrinsic image evaluation. This dataset uses synthetic data and contains both simple objects and complex scenes under different illumination settings. We show that it is possible to easily create large and realistic datasets for intrinsic image evaluation using computer graphics software and rendering engines.

Finally, in Chapter 6 we list our contributions and discuss future work about how the lines of research presented in this thesis could be further developed.

# Chapter 2

## Review on Computational Intrinsic Images

Intrinsic images were first introduced by Barrow and Tenenbaum in [23]. The authors defined an intrinsic image of a given scene as an image depicting a single physical characteristic of the scene such as reflectance, illumination, orientation, distance to the observer, transparency, specularity, luminosity, etc. Figure 2.1 illustrates how the authors imagined some of these components (top row) and their actual representations (bottom row). Intrinsic image decomposition is a hard and ill-posed problem, because the number of unknowns is greater than the number of equations. This means that multiple combinations of different intrinsic components may result in the same input image. However, human beings have the ability to interpret images, and recognize and isolate specific physical properties of the scene. For example, we can determine the actual color of the objects in a scene, even if their appearance is modified by an illuminant. We are also able to interpret the shape of an object or the depth in a scene. The goal of intrinsic image decomposition approaches is to computationally emulate the human ability to isolate the different factors that form an image. This chapter provides a detailed and organized overview on these methods for intrinsic image estimation.

Intrinsic images describe specific features that provide a better understanding of scenes and facilitate subsequent processing. Intrinsic images have been widely used in many subfields of computer vision, such as shape from shading [45], color constancy [63], highlight removal [13] or color photo editing [108, 25]. However, the so-called intrinsic image decomposition methods initially focused on providing reflectance and shading image estimates of a scene.

In order to perform such a decomposition, early methods focused on making simplifying assumptions about the scenes and classifying image derivatives based on the Retinex theory [104, 103]. Reformulating the intrinsic image decomposition as an energy function optimization problem allowed the inclusion of multiple and diverse information cues into the equation, which resulted in better reflectance and shading estimates. Recently, the development of technology has enabled methods in intrinsic image decomposition to include extra input information, which simplifies the estima-

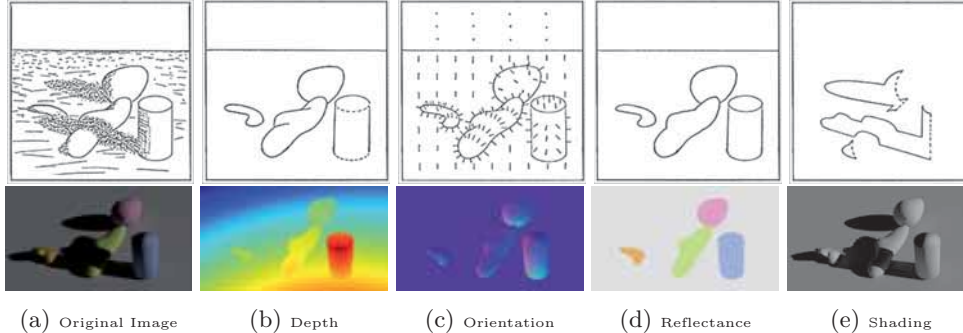


Figure 2.1: Intrinsic components as illustrated by Barrow and Tenenbaum in [23] are shown in the top row, while actual representations of these intrinsic components can be found in the bottom row. Representations for depth and orientation in [23] are based on the 2.5D sketch of Marr [117].

tion process. A good example of this is the appearance of RGB-D cameras such as Kinect [39]. Such cameras allowed the consideration of coarse information about the depth of the scene.

Nevertheless, the mechanisms for intrinsic image evaluation have not evolved accordingly. The appearance of the MIT dataset [71], which provided ground truth data for reflectance and shading components as well as specularities, was an important step towards the objective evaluation and comparison of intrinsic image decomposition methods. However, recent methods make it necessary to build larger datasets which include ground truth data for multiple intrinsic characteristics of the scene.

Although a lot of progress has been made in the field of intrinsic images during the last decade, to the best of our knowledge no exhaustive review of the field is available in the literature. The objective of this chapter is to present a comprehensive and organized review of the techniques for intrinsic image estimation. Methods on intrinsic image decomposition can be reviewed in many different ways. In this work, we use different principles to review the field. The first one, in Section 2.1, is based on the different assumptions that the methods for intrinsic image decomposition have made about the world in order to simplify the problem of intrinsic image decomposition. This principle gives us insight into the theoretical modeling underlying the problem. In Section 2.2, we introduce different visual cues based on regularities in the scenes and images which have been used in different intrinsic image estimation methods. Some of these visual cues have been fundamental in the field and have influenced most works on intrinsic image decomposition. The other principle, in Section 2.3, focuses on practical aspects of the problem such as the number and sort of input images that each method uses, the number and sort of intrinsic components being estimated, the nature of the information cues used to constrain the solution (introduced in Section 2.2) and the techniques used to solve the problem. Although assumptions are used in all these sections, they refer to different kinds of constraints. While the assumptions described in Section 2.1 refer to general regularities about the physics of the scenes and have an influence on the formulation of the problem, the assumptions presented in Sections 2.2 and 2.3 are focused on regularities that affect the image formation process. In



Section 2.4, we also review the evaluation mechanisms that have been used so far in the field of intrinsic image decomposition, analyzing the existing datasets and metrics. Finally, we report the conclusions of this survey and analyze recent trends in the field and potential developments in Section 2.5.

## 2.1 Problem formulation and simplifying assumptions about the scenes

As mentioned above, the decomposition of an image into its intrinsic components is an ill-posed problem with more unknowns than equations. In order to simplify the problem, all approaches make simplifying assumptions about the complexity of the world. In this section we describe the most common assumptions and how they constraint the theoretical formulation of the problem.

Assuming Lambertian surfaces in a scene is one of the most widely used hypothesis in intrinsic image decomposition. The luminance of Lambertian surfaces is isotropic, which means that there are no specularities\*. Under this assumption, many methods have been built to decompose the input image into its shading and reflectance components,

$$I(x, y) = I_{Shad}(x, y) \cdot I_{Refl}(x, y), \quad (2.1)$$

where  $I_{Shad}$  is the shading image and represents the amount of reflection of the light arriving at the point  $(x, y)$  of the image from a specific point of the object surface considering the shape of the objects in the scene and the position of the light source, and  $I_{Refl}$  is the reflectance image, which describes how the light is reflected by the corresponding point of the object considering the material reflectance properties.

The form of the intrinsic components also depends on the assumptions we make about the world.  $I_{Refl}$  has been usually represented as a 3-channel matrix, where each channel represents one of the dimensions of the RGB color space. When the world is assumed to be achromatic and the input image lacks color information,  $I_{Refl}$  is a grey-scale image and is represented as a single channel matrix [141, 27, 157, 145, 20]. By contrast,  $I_{Shad}$  has been usually represented as a 1-channel matrix. When it is the case, the shading models the local intensity of the reflectance at each pixel or, in other words, scenes are assumed to be illuminated by a white light (*i.e.* an achromatic light source that affects the intensity of the reflectances in the scene but not their chromaticity). However, a few methods have assumed more complex worlds and use 3-channel matrices to describe the shading images, allowing to model color illuminants and/or interreflections in the scene [32, 102, 19, 21]. The shading image,  $I_{Shad}$ , has also been generalized as a function,  $M()$ , of the shape of the objects in the scene,  $I_{Shape}$ , and the model of illumination,  $\mathbf{L}$ , in different works [20, 19, 105, 102, 144, 21, 38, 154], leading to the following formulation of the problem

$$I(x, y) = I_{Shad}(x, y) \cdot I_{Refl}(x, y) = M(I_{Shape}(x, y), \mathbf{L}) \cdot I_{Refl}(x, y). \quad (2.2)$$

---

\*Specular reflections of the scene, which often result in saturated values in the images.

When the Lambertian assumption is relaxed, methods must cope with potential specularities in the input images [143, 42, 105, 21]. Some of these works [105, 21] have included the highlights in the shading image, while the others [143, 42] have specifically modeled these effects and have included them in the formulation of the problem,

$$I(x, y) = I_{Shad}(x, y) \cdot I_{Ref}(x, y) + I_{Spec}(x, y), \quad (2.3)$$

where  $I_{Spec}$  denotes the specular reflection of the objects in the scene and acts as an additive term.

Although both the color of the illuminant and the camera sensors influence the pixel values we observe in the input images, none of the previous formulations model their effects. In Chapter 4 we present a new framework for intrinsic image decomposition which models the effects of the illuminant of the scene and the camera sensors. Our framework generalizes the formulations described in this section.

### 2.1.1 Discussion

Adding new terms to the basic formulation of the problem (Equation 2.1) without including extra equations may not seem a good idea because we are underconstraining the problem. However, these terms are closely related and allow us to use prior knowledge about the world in order to constrain the number of possible solutions and determine the most likely ones. A good example of this idea is illustrated in Figure 1.9, where different combinations of shape, lighting, and reflectance result in the same image. However, according to a set of predefined independent rules about the world (*i.e.* prior knowledge), there is one explanation of the intrinsic composition of the image which is more probable than the others. Table 2.1 summarizes how the existing methods have used the simplifying assumptions reviewed in this section.

## 2.2 Visual cues based on scene and image regularities

In this section we review different visual cues that have been used in intrinsic image estimation. These cues are based on physical regularities about the images and the underlying scenes and so far have proved to be extremely useful in order to bound the space of plausible solutions to the problem of intrinsic image decomposition. Typical regularities are that the curvature of most objects in a scene usually changes smoothly, or that the reflectances in a scene are usually sparse and can therefore be represented in reflectance images with a reduced set of color values.

### 2.2.1 Shading Smoothness

The first attempt to divide the lightness of an image into its reflectance and shading components (Equation 2.1) is found in the Retinex theory [104, 103], which was released before the intrinsic image problem was formally defined in 1978. The Retinex aimed to emulate some retinal-cortical processes in the human visual system in a 2D

Work	Lambertian Surfaces	Single Uniform Illuminant	White Light
Funt'92 [57]	✓	✓	✓
Sinha'93 [141]	✓	✓	✓*
Bell'01 [27]	✓	✓	✓*
Weiss'01 [157]	✓		✓*
Finalyson'04 [49]	✓	✓	
Matsushita'04 [119]	✓		✓
Olmos'04 [122]	✓	✓	✓
Tan'04 [143]		✓	
Tappen'05 [146]	✓	✓	✓
Tappen'06 [145]	✓	✓	✓*
Bousseau'09 [32]	✓	✓	
Jiang'10 [90]	✓	✓	✓
JShen'11 [136]	✓	✓	✓
LShen'11 [138]	✓	✓	✓
Gehler'11 [60]	✓	✓	✓
Barron'12a [20]	✓		✓*
Barron'12b [19]	✓		
Garces'12 [59]	✓	✓	✓
Lee'12 [105]	✓†		✓
Serra'12 [133]	✓	✓	✓
Laffont'12 [102]	✓		
Tang'12 [144]	✓	✓	✓
Zhao'12 [137, 161]	✓	✓	✓
Barron'13 [21]	✓†		
Chen'13 [38]	✓	✓	
Vineet'13 [154]	✓		
Chang'14 [37]	✓	✓	
Jeon'14 [89]	✓	✓	✓
Ye'14 [159]			✓
Kong'14 [99]			✓
Bell'14 [28]	✓	✓	✓

Table 2.1: Classification of the methods based on the simplifying assumptions they make about the scenes. † Specularities are included in the shading image. \* Grayscale images are used.

Mondrian world<sup>†</sup> of Lambertian surfaces. These processes are thought to be at the basis of the human ability to perceive color robustly and independently of the color of the illuminant (color constancy).

The authors showed in their experiments that adjacent luminances do not differ abruptly unless there is a boundary between two areas with distinct reflectances. Therefore, luminance edges in images are an important source of information: brusque changes indicate reflectance variations while smooth variations are due to shading. Using these ratios of luminance at adjacent points, the Retinex algorithm can both detect edges and eliminate the effect of nonuniform illumination or shading.

However, the assumptions of the Retinex do not always hold when decomposing natural images, which are more complex than 2D Mondrian worlds. The 3D geometry of a scene can produce sharp shading edges when occluding objects cause cast shadows, and the materials of the objects, which are generally non-Lambertian, often result in specularities.

Although sharp shading edges can be found in natural images, shading variations are still smooth in most of the images. As a consequence, the Retinex-like classification of image derivatives into those caused by reflectance changes and those caused by shading variations has been successfully and recurrently used in many successive works [57, 141, 27, 157, 119, 49, 122, 143, 146, 145]. Moreover, many of the works that do not classify image derivatives [138, 60, 133, 59, 161, 89] also include different cues based on the assumption that shading varies smoothly as proposed in the Retinex theory.

In [57], Funt *et al.* observed that chromaticity edges could be useful to detect luminance variations caused by sharp shading edges, since shading variations are theoretically cancelled in chromaticity images. This property has been used in most posterior works also including shading smoothness cues.

Other existing cues are also based on this shading smoothness regularity. In [19], Barron and Malik defined a cue on the smooth variation of mean curvature in the shape of the objects. In fact, shading smoothness is a direct consequence of curvature smoothness. Following this idea, Lee *et al.*, in [105], proposed that non-neighboring pixels with the same surface normal direction should have similar shading descriptions.

## 2.2.2 Texture Structure

Regularities on texture have been used as well to estimate intrinsic images. Textures can be caused by both reflectance and geometric patterns. In [106], Leung and Malik defined a vocabulary of surface patches with associated geometric and photometric properties called 3D textons. These 3D textons have been used in some intrinsic image decomposition works to define a visual cue based on texture structure [137, 105]. In these works, pixels sharing similar local texture structures were grouped together as having the same reflectance values.

---

<sup>†</sup>A world made of color patches

### 2.2.3 Color Sparsity

The number and distribution of color in scenes is another important regularity which has been exploited in the problem of intrinsic image estimation. Multiple factors affect the color values we observe in images. Some color distortions are produced by the geometry of the objects in the scene and the nature of the light sources (highlights, shadows, interreflections, etc.). Other factors influence the color values during the image acquisition process (camera sensors, non-linear transformations, etc.). All these issues, the ones that happen in the scene itself and the ones that occur during the image acquisition process, cause that pixels which represent the same reflectance are finally described with different color values in images.

Omer and Werman, in [123], analyzed the distribution of image reflectances in color histograms and showed that the reflectances of natural scenes are usually sparse. Therefore, these reflectances can be efficiently described by a sparse and reduced set of color values in the reflectance intrinsic images. The authors introduced a method which clusters image pixels based on the analysis of reflectance distributions in color histograms. In a similar fashion, in [151], Vazquez *et al.* presented a method which analyzes the local maxima of a color distribution in a color histogram space in order to cluster image pixels. These methods provide a compact description of reflectances which is robust to color distortions.

Color sparsity is also a central concept in human perception. Although humans can distinguish millions of different colors [91], it has been studied that human beings from around the world use a very discrete set of semantic terms to describe perceived colors [31].

Using a sparse set of colors as a cue has been recurrent in the definition of energy functions for intrinsic image estimation [32, 138, 60, 133, 59, 19, 28]. This idea is simply to enforce neighboring pixels with a similar chromaticity to share the same reflectance value. Assuming that the reflectances in a scene can be described with a sparse set of colors allows us to reduce the number of possible solutions, but it can be a problem when representing scenes containing many different reflectances or smooth reflectance changes.

### 2.2.4 Discussion

Assumptions about regularities in scenes and images reduce the space of potential solutions of the intrinsic image decomposition problem. The inclusion of new intrinsic components into the formulation of the problem usually implies the use of new assumptions. The color sparsity cue, for example, was only considered when methods started using input color images. The visual cues described in this section will be used as a classification factor in the next section.

The actual use of RGB-D images and video, has also enforced the definition of new cues based on shape and temporal regularities. Some of these cues will be mentioned in the following section.

Work	Scene and Image Regularities		
	Shading Smoothness	Color Sparsity	Texture Structure
Funt'92 [57]	✓		
Sinha'93 [141]	✓		
Bell'01 [27]	✓		
Weiss'01 [157]	✓		
Finalyson'04 [49]	✓		
Matsushita'04 [119]	✓		
Olmos'04 [122]	✓		
Tan'04 [143]	✓		
Tappen'05 [146]	✓		
Tappen'06 [145]	✓		
Bousseau'09 [32]	✓	✓	
Jiang'10 [90]	✓		✓
JShen'11 [136]	✓	✓	
LShen'11 [138]	✓	✓	
Gehler'11 [60]	✓	✓	
Barron'12a [20]	✓	✓	
Barron'12b [19]	✓	✓	
Garces'12 [59]	✓	✓	
Lee'12 [105]	✓		✓
Serra'12 [133]	✓	✓	
Laffont'12 [102]	✓		
Tang'12 [144]			
Zhao'12 [137, 161]	✓		✓
Barron'13 [21]	✓	✓	
Chen'13 [38]	✓		
Vineet'13 [154]	✓	✓	
Chang'14 [37]	✓	✓	
Jeon'14 [89]	✓		✓
Ye'14 [159]	✓		
Kong'14 [99]	✓	✓	
Bell'14 [28]	✓	✓	

Table 2.2: Classification of the methods according to their assumptions about the most common regularities in the scenes.

## 2.3 Intrinsic image decomposition techniques

In this section we propose a 2-layer classification of methods for intrinsic image decomposition based on the estimation techniques they use, as well as the information cues they use and their input and output information. Estimation techniques provide the first criterion of the classification. At this level, the methods are divided between these approaches which focus on labeling image derivatives and the others, which mostly formulate the estimation of intrinsic components as an energy function optimization problem. Information cues used to constrain the problem, as well as the input information the approaches require, are used in the second level of this classification. Some of these cues have been described in the previous section. Input information mainly refers to the number (single *vs.* multiple) and sort (grayscale, color, RGB-D, etc.) of images that each method requires, but also considers other sources of input information such as user assistance. Although some approaches fulfill more than one criterion and could be included in more than one category, we classify them according to their most relevant characteristics.

### 2.3.1 Classification of image derivatives

Several methods have attempted to extend the Retinex theory, introduced in Section 2.2, to analyze images representing a more realistic world. For example, Sinha and Adelson, proposed in [141] the classification of image edges in a 3D achromatic world of painted polyhedra without object occlusions and cast shadows. In this theoretical work, intensity junctions (*i.e.* intensity edge intersections) were locally analyzed and classified as being caused by either shading or reflectance variations. A post-processing step to verify the consistency of the 3D structure of polyhedra and the illumination source direction was also proposed.

Other methods have used information from different sources in order to achieve a better classification of image derivatives.

#### Multiple images

In an attempt to simplify the problem of derivative classification, Weiss, in [157], used multiple images as an input. These images shared the same reflectance under different illumination conditions. In such framework, shading edges were different throughout the images while reflectance variations were preserved. Derivative filters were applied to the different frames, and a median of these filtered image outputs was used to avoid luminance edges caused by shading variations and recover an intrinsic reflectance image. An extension of this method was proposed by Matsushita *et al.* in [118, 119], where a threshold was used to remove edges caused by texture patterns from the illumination images. In this work, the authors also proposed to build an illumination eigenspace which allowed real-time estimation of shading images in fixed scenarios, such as these provided by traffic cameras.

### Chromaticity edges

In a different direction but still using the same fundamental principle of classifying image derivatives, some other works have exploited color information. Funt *et al.*, in [57], adapted the Retinex theory to classify chromaticity derivatives instead of luminance edges. They claimed that in chromaticity images shading variations are canceled, while reflectance differences are preserved. Chromaticity edges have been used in other works. For example, Olmos and Kingdom assumed in [122] that co-aligned chromatic and luminance variations usually describe changes in surface reflectance in natural scenes, whereas other luminance variations mainly arise from shading and shadows. Finlayson *et al.* also combined chromatic and luminance derivatives in [49] to detect and remove shadow edges and recover shadow-free images, which could be further decomposed into their shading and reflectance components. Chromaticity edges also played a fundamental role in [143], where Tan and Ikeuchi first generated a specular-free image by shifting the intensity and chromaticity of pixels while retaining their hue. Then, reflectance derivatives were removed from this specular-free image, and the resulting shading image was used to decompose the reflection components of the input image into its diffuse and specular reflections. The diffuse component was used to calculate the reflectance image.

### 2.3.2 Learning-based approaches

Instead of just classifying image derivatives using a simple threshold, some authors have proposed approaches based on learning. For instance, Tappen *et al.* [146] trained a classifier using the AdaBoost algorithm [56] to recognize gray-scale intensity patterns. They combined this classifier with information about chromatic edges in order to label image derivatives. Global spatial coherence for the edge labeling was achieved in a last stage using a Markov random field solved using the Generalized Belief Propagation algorithm [160]. The same authors, in [145], used a training set composed of images of real surfaces to build estimators that predicted local shading and reflectance derivatives from image patches. Then, they learned a weighting function that weighted the different local estimates in order to produce the best possible global estimate. Bell and Freeman, in [27], used a training set of synthetic images containing both shading and reflectance variations and learned a classifier for steerable pyramid coefficients [140]. These coefficients represented the outputs of first and second derivative filters which had been applied to the luminance of the input image. A final propagation stage provided global coherence to the method. Recently, Tang *et al.*, in [144], combined deep belief networks [78] with the Lambertian reflectance assumption in order to learn priors on the albedo from images and use this knowledge to estimate the albedo and surface normals of similar images.

### 2.3.3 Energy functions optimization

Most of the works which have not focused on classifying image derivatives have formulated the decomposition problem as the optimization problem of an energy function. In these works, the authors first define a function encoding all the assumptions about



the problem and then select an optimization technique able to converge to a global minimum of the given energy function. So far, these optimization approaches have proved to achieve satisfactory results, as detailed below.

In order to reduce the space of solutions, information from multiple sources has been used in these optimization approaches.

### Sparsity color cue

The use of a global sparsity color cue [123, 151], whose basic idea is that the reflectances of natural images can be described by a sparse and reduced set of color values, has been recurrent in the definition of energy functions for intrinsic image estimation. Shen and Yeo defined in [138] an energy function that combined the global reflectance sparsity constraint with a local Retinex cue on chromaticity edges, which enforced neighboring pixel with similar chromaticities to share the same reflectance values. The energy function was optimized using least-squares minimization [94]. Gehler *et al.* defined, in [60], a probabilistic model where this global sparsity prior on reflectance was combined with a color Retinex cue which extracted potential reflectance edges, and an intensity Retinex cue which enforced shading smoothness between neighboring image pixels. To optimize the energy function, which was expressed as a latent variable random field, the authors used a coordinate descent algorithm [112]. Garces *et al.*, in [59], divided the image into clusters of similar chromaticity using this global sparsity cue and used a linear system to enforce shading smoothness on the boundaries between clusters. The authors used a Quasi-Minimal Residual method [18] to solve the system. Bell *et al.*, in [28], combined different priors from previous works, including Retinex cues on intensity and chromaticity edges and the global sparsity cue in a conditional random field and solved the inference problem using the method presented in [100].

Although most energy optimization methods presented in this section are discriminative (*i.e.* they model the dependence of unobserved variables from observations), Chang *et al.*, in [37], presented a Bayesian generative model based on the work of Gehler *et al.*. This work overcomes many similar discriminative methods without including any Retinex-like term in their problem formulation. The authors used Markov Chain Monte Carlo techniques [73] to solve their probabilistic non-parametric approach.

### Texture cues

Still using energy functions, some authors have also considered texture cues for solving the decomposition problem. For instance, Shen *et al.* [137] searched for pixels in the image which shared similar local texture structures [106] in the chromaticity image and grouped these pixels together as having the same reflectance. These texture constraints were combined with local color Retinex constraints. To optimize the energy function, the authors expressed it in a graph structure and used tree-reweighted message passing [98]. This work was extended in [161], where Zhao *et al.* reformulated the problem as the minimization of a quadratic function and used the standard conjugate gradient algorithm [77] to solve it. In a different fashion, Jiang *et al.*, in [90], separated images into frequency and orientation components using steerable

filters [140] and constructed shading and reflectance images from weighted combinations of these components. These weights were determined by correlations between corresponding variations in local luminance (*i.e.* mean of intensities), local amplitude (*i.e.* variance of intensities), color and texture, assuming that positive correlations between luminance and color or texture usually indicate reflectance changes, while correlated changes in mean luminance and luminance amplitude are probably due to illumination variations.

### User-assistance

In the work of Bousseau *et al.* [32], the users used sparse strokes to mark parts of the image that shared the same reflectance and parts where illumination did not vary. Users were also expected to provide at least one fixed illumination value to solve the global scale ambiguity. The authors built a constrained least-square system and used a multigrid solver [34] to optimize an energy function over local windows whose only variable was the shading. This energy function combined the constraints provided by the users with a global sparsity color cue [123]. The reflectance image was directly recovered dividing the original image by the estimated shading image. User strokes were used again in [136], where Shen *et al.* proposed an automatic optimization algorithm based on the assumption that neighboring image pixels with similar intensity values share similar reflectance values. Information about user strokes was used to constrain the minimization of the energy function, but the optimization technique the authors used to minimize the energy function was not specified.

### Shape cues

Recently, some energy optimization approaches have included additional outputs in the formulation of the problem. In [20], Barron and Malik defined a probabilistic framework where shading, reflectance, shape and an illumination model were jointly estimated from a single gray-scale image and its corresponding mask. They designed an energy function which contained multiple priors on reflectance and shape. The priors on reflectance were based on the global sparsity color cue and the local Retinex cue about reflectance edges. The priors on shape included a cue on flatness enforcement to address the bas-relief ambiguity [26], another one on occluding contours, which assumed input masked scenes representing single objects with known orientation at boundaries, and finally one cue on the smooth variation of mean curvature. When the illumination was unknown, it was estimated from a discrete set of spherical harmonic illuminants. This method was extended in [19] to color input images and was called SIRFS (shape, illumination and reflectance from shading). The priors on shape were kept while the priors on reflectance were adapted to color images and a new cue on color-constancy, which enhances reflectances that are close to white or that lie within the gamut of previously seen colors, was added. The illumination was included as a prior, and they used a multivariate Gaussian model which had been fit to their training set of spherical harmonic illuminations. The authors used a L-BFGS technique [36] to minimize their energy function.

### RGB-D images

The emergence of new devices such as Kinect cameras, which provide RGB-D images (*i.e.* color images that include a depth map), has allowed multiple intrinsic decomposition methods to use this shape information as an input. Lee *et al.*, in [105], used multiple RGB-D images to estimate reflectance and shading intrinsic components. The authors defined an energy function based on different sets of constraints and minimized it using a sparse linear solver [74]. They used reflectance constraints based on the Retinex-like cue on chromaticity edges and the non-local texture constraints defined in [137]. They also proposed shading constraints based on shading smoothness not only between neighboring pixels, but also between pixels that share the same surface normal direction. Temporal constraints were proposed as well to reduce the effects of noise, using the average color among the corresponding points in the different frames, and to determine intensity outliers in temporally distant frames and avoid specularities. Barron and Malik, in [21], used a single RGB-D image to solve a probabilistic framework based on their previous work [19]. The authors used the depth input as an observation and removed their prior on occluding contours, allowing the model to work for unmasked scenes. Chen and Koltun [38] proposed a linear least squares formulation [94] of the problem where they estimated the reflectance component and further decomposed the shading component into a direct irradiance component, an indirect irradiance component (*i.e.* interreflections), and a color component (*i.e.* color of the illuminant). They defined an energy function based on chromaticity edges and their observations on irradiance. The authors stated that direct irradiance varies slowly as a function of position and surface orientation while indirect irradiance can have higher frequencies, and employed regularizers that model these characteristics. Jeon *et al.*, in [89], first decomposed an input RGB-D image into a texture layer, containing the texture patterns in the image, and a base layer, where these texture patterns had been removed. Then, this base layer was further decomposed into its reflectance and shading components, assuming Retinex-based constraints and other local and global shading constraints based on surface normals. The authors used a conjugate gradient method [77] to minimize the energy function. In a different direction, Vineet *et al.*, in [154], jointly estimated intrinsic image components and semantic properties about objects and material attributes of the scene using a high order conditional random field. Their model used the method defined in [19] and the high-order potentials they proposed were based on the correlations between the reflectance, objects, and attribute labels assigned to the pixels. In order to optimize their energy function, the authors used approximate dual decomposition [148].

### Photo collections

The appearance of photo collections on the internet has also been used for intrinsic image decomposition. Some methods have used this extra information to simplify the optimization process of the energy function. In [102], Laffont *et al.* used multiple input images of the same scene from internet photo collections in order to infer object geometry using patch-based multi-view stereo [58] and object reflectance based on Weiss' method [157]. The authors used a blockwise Gauss-Seidel solver [87] to

optimize their sparse linear system. Further applications of their method, such as relighting of scenes and color transfer, were also presented in this work.

## Video

Video has been used as well for intrinsic image decomposition [105, 159, 99]. Although the different frames of a video can be seen as the multiple images used in previous methods [102, 157], there is an important difference. Methods using video as an input do not assume a rigid scene structure as it happens with photo collections [102]. Methods based on video do not assume either that the camera and the objects in the scene are fixed and the light changes, as in [157]. In a video the different objects in the scene can be moved independently. Lee *et al.*, in [105], only used the video information to include temporal constraints in their energy function, which reduced the effects of different artifacts such as noise or specularities. Ye *et al.*, in [159], defined a Maximum a Posteriori problem. The authors first decomposed the first frame of the video into its shading and reflectance components. Then, reflectance values were propagated on successive frames until a confidence threshold on the number of unassigned pixels stopped this propagation. Local decomposition was performed on the unassigned pixels at the stopping frame and the new values were propagated backward in time. Finally, a smoothness constraint was applied on the shading of the few remaining unassigned pixels. Kong *et al.*, in [99], also took advantage of video frames to jointly estimate optical flow and intrinsic albedo and shading components. Optical flow information was used to enhance temporal constancy for reflectance and temporal smoothness for shading. The authors used a coarse-to-fine pyramid-based approach based on the Classic+NL flow estimation method in [142] to minimize their energy function.

### 2.3.4 Discussion

In this section we have provided a classification of the methods based on their estimation technique (first layer of the classification) and their input information (second layer of the classification). The assumptions about the regularities in the scenes explained in the previous section have also been used in this classification. We have observed that methods on intrinsic image decomposition have evolved in parallel with the development of both technology and optimization techniques. Initially, most methods, based on the Retinex theory [103], focused on classifying image derivatives. However, most current methods pose the decomposition problem as an optimization problem. On the one hand, the appearance of powerful optimization techniques which are able to minimize complex energy functions has allowed recent models to jointly estimate multiple intrinsic components other than reflectance and shading. On the other hand, the appearance of RGB-D cameras or big internet image collections has provided extra input information which has proved to be very useful. In the next section we will see how this evolution of the methods has surpassed the existing evaluation mechanisms.

The information presented in this section has been synthesized in Tables 2.3 and 2.2. Output information about the intrinsic components that each method estimates

Work	Inputs			Technique			Outputs		
	Color	Depth	Mult. Im.	User Ass.	Edge Class.	Energy Opt.	Ref. & Shad.	Shape	Others
Funt'92 [57]	✓				✓	✓	✓		
Sinha'93 [141]					✓		✓*		
Bell'01 [27]					✓	✓	✓*		
Weiss'01 [157]			✓		✓		✓*		
Finalyson'04 [49]	✓				✓	✓	✓		(1)
Matsushita'04 [119]			✓		✓		✓		
Olmos'04 [122]	✓				✓		✓		
Tan'04 [143]	✓				✓		✓		(2)
Tappen'05 [146]	✓				✓	✓	✓		
Tappen'06 [145]					✓		✓*		(3)
Bousseau'09 [32]	✓			✓	✓	✓	✓†		
Jiang'10 [90]	✓						✓		
JShen'11 [136]	✓			✓		✓	✓		
LShen'11 [138]	✓					✓	✓		
Gehler'11 [60]	✓					✓	✓		
Barron'12a [20]						✓	✓*	✓	(4)
Barron'12b [19]	✓					✓	✓†	✓	(4)
Garces'12 [59]	✓					✓	✓		
Lee'12 [105]	✓	✓				✓	✓		
Serra'12 [133]	✓					✓	✓		
Laffont'12 [102]	✓		✓			✓	✓		
Tang'12 [144]							✓	✓	
Zhao'12 [137, 161]	✓					✓	✓		
Barron'13 [21]	✓	✓				✓	✓†	✓	(4)
Chen'13 [38]	✓	✓				✓	✓		(5)
Vineet'13 [154]	✓	✓				✓	✓	✓	(6)
Chang'14 [37]	✓					✓	✓		
Jeon'14 [89]	✓	✓			✓	✓	✓		(7)
Ye'14 [159]	✓		✓	✓	✓	✓	✓		
Kong'14 [99]	✓		✓			✓	✓		(8)
Bell'14 [28]	✓					✓	✓	✓	

Table 2.3: Classification of the methods based on their decomposition technique as well as their inputs. The outputs of the methods have also been included in this table for the sake of completion. Other outputs: (1) Isolated shadow information. (2) Specular component. (3) Isolated noise effects. (4) Model of illumination. (5) Direct and indirect irradiance information. (6) Object and attribute labels. (7) Texture component. (8) Optical Flow. \* Grey-scale reflectance images. † Color shading images.

has also been included in the Table 2.3 for the sake of completion.

## 2.4 Intrinsic Image Evaluation

Being able to objectively evaluate and compare the performances of different existing methods on a given problem is essential in any engineering field. However, early works on intrinsic image decomposition could not provide quantitative results, since no ground truth dataset existed and no metric was specifically defined to evaluate the accuracy of intrinsic images. In this section we describe the different datasets and metrics that have been used so far to evaluate intrinsic images.

### 2.4.1 Qualitative examples

The quantitative evaluation of methods for intrinsic image decomposition has so far proved difficult. Although some datasets exist, many authors still provide qualitative results to assess the performance of their methods, specially when they want to demonstrate the benefits of their approaches in a specific type of images. Therefore, we find it interesting to briefly mention in this section some of the images and datasets that have been used for qualitatively testing multiple intrinsic image decomposition methods.

Weiss, in [157], presented a method which takes a set of images of a single scene under different lighting conditions as its input. The author provided qualitative results for webcam images and images from the Yale Face database B [61], which contains thousands of facial images of 28 subjects taken under different poses and illumination conditions and has also been used in [144]. In a similar fashion, Matsushita *et al.*, in [118, 119], used real sequences of road traffic images.

Some authors have provided qualitative results for complex images representing diverse natural scenes [122, 146, 32]. In [90], Jiang *et al.* presented results for a subset of images from their own dataset, the Birmingham Object Lighting Database, which contains stereoscopic image pairs of objects, surfaces, faces and outdoor scenes photographed with high resolution under well specified lighting conditions. Other methods dealing with specific effects, such as cast-shadows [49], specularities [143] or textures [137], used their own test images for visual evaluation (see some examples in Figure 2.2).

Internet photo collections, such as Flickr, have also been an important source of images for visually testing intrinsic image methods [137, 32, 59, 102]. Shen *et al.*, in [137], used images containing textures, while Laffont *et al.*, in [102], looked for multiple pictures of different famous landmarks. Bousseau *et al.* [32] provided qualitative examples on many different and diverse scenes, some of which were used in further works [138, 59, 133].

The appearance of new methods which use RGB-D images as inputs and estimate other intrinsic scene components in addition to shading and reflectance, forced recent methods to use more complex datasets for testing. The NYU Depth dataset [139] contains video sequences of 464 different indoor scenes recorded by both the RGB and depth cameras from the Kinect, and 1449 densely labeled pairs of aligned RGB and depth images. Although it was not specifically designed for intrinsic image evaluation,

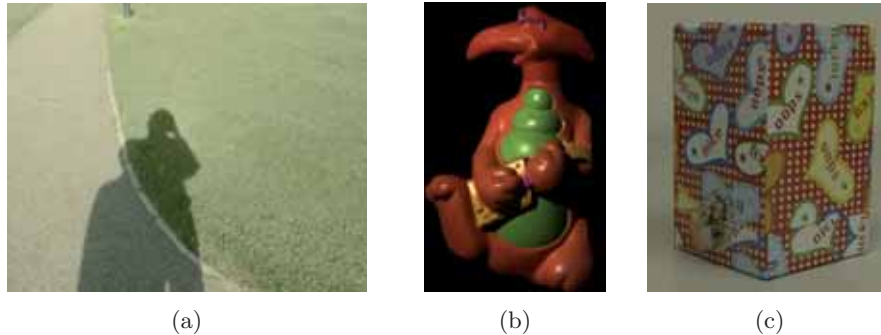


Figure 2.2: Input images for different methods. (a) Image containing cast shadows [49]. (b) Image with specularities [143]. (c) Highly textured images [137].

it provides useful intrinsic information for the decomposition problem, such as depth maps, and has been used for qualitative evaluation in different works [21, 154, 38, 89].

## 2.4.2 Datasets

In order to evaluate intrinsic image estimates, we need ground truth data for the different intrinsic components. Obtaining these ground truth data is a challenging problem, since isolating intrinsic components of natural scenes is not easy. In this section we discuss the different datasets that have been used to test the existing methods.

### Early ideas

The first idea of a ground truth image for intrinsic image evaluation is found in [27], where Bell and Freeman showed two images of a white wall: one of the wall with a graffiti, and another after the wall was repainted white (Figure 2.3). Assuming that no reflectance variations occur in this second image, it can be thought of as an intrinsic shading image of the scene.

Tappen *et al.*, in [145], provided the first quantitative evaluation and the first ground truth dataset for the intrinsic image decomposition problem. The dataset consists of 46 images of wrinkled papers which have been colored with a green Crayola marker. The authors used the green channel of the images, where these markings are not visible, as ground truth shading. The reflectance images were obtained by means of a simple point-wise division. An example of the Crayola dataset is shown in figure 2.4. The main problem of this dataset is that all its images are very similar, making it difficult to determine if a method which performs well with this dataset will also achieve good results with any other sort of images.

### The MIT dataset

The release of the MIT dataset for intrinsic image evaluation [71] was an important step towards the comparison of the different existing algorithms. The authors built



Figure 2.3: Graffiti image from [27]. (a) Graffiti painted on a wall. (b) Graffiti covered with white paint.

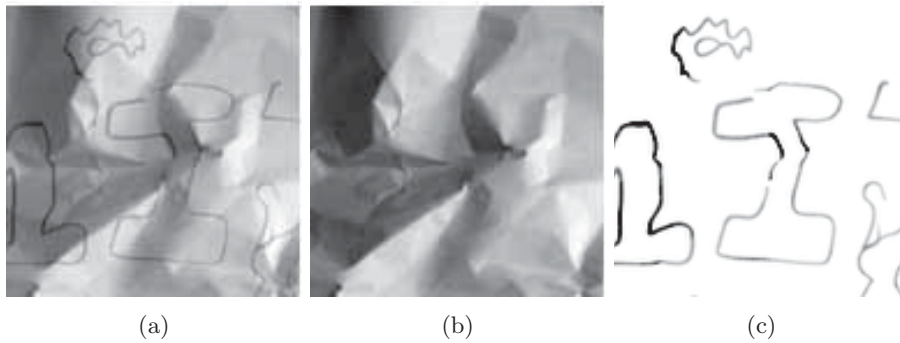


Figure 2.4: Sample image from the Crayola database [145]. (a) Original image: wrinkled paper colored with a Crayola marker. (b) Shading image: the Crayola marker is invisible in the Green channel of the image. (c) Reflectance image: result of dividing the original image by the shading image.

a dataset containing images of 16 toy-like objects (although 4 extra objects can be found in their website). The objects in the dataset are shown in Figure 2.5. For each object the following data are supplied: the original image, a diffuse image obtained using a polarizing filter, a binary mask of the object, a shading image which results from painting the object white, a reflectance image, and a specular image which is the result of subtracting the diffuse image from the original. The authors also provided the original images under multiple illumination settings. These images are necessary for methods which require multiple input images such as [157, 119].

Building such a dataset was challenging. In order to avoid complex and undesired effects such as interreflections, highlights and transparencies, the shapes of the objects are essentially convex, and their materials are mostly diffuse (the few remaining specularities were removed with a polarizing filter). Moreover, the method used for capturing the shading images of an object consisted in carefully removing the object, painting it white and replacing it at the exact same place. This method is costly in time and hardly applicable to natural scenes, which explains the reduced number and



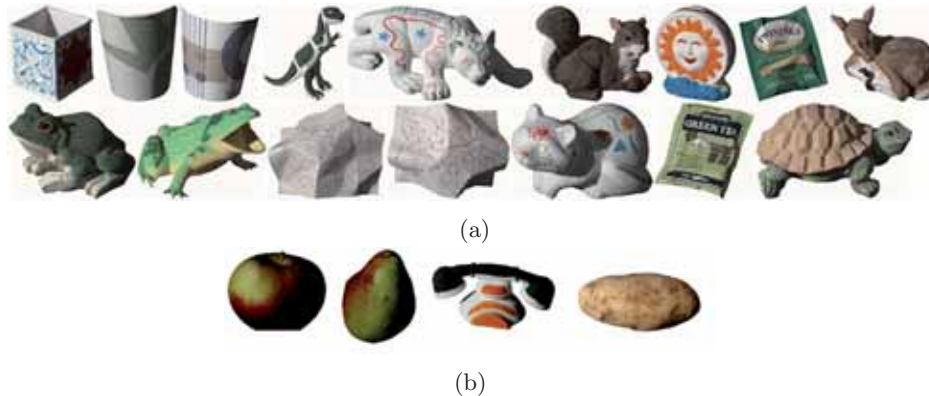


Figure 2.5: Images in the MIT dataset [71]. (a) The 16 original objects that composed the dataset. (b) The 4 extra objects that can be found on the authors' website.

the small variety of objects found in the MIT dataset.

Nonetheless, the MIT dataset [71] was the first satisfactory proposal of a ground truth dataset for intrinsic image evaluation. Since its release, most authors working on estimating reflectance and shading images have used this dataset to test their methods and compare them with other approaches. Some methods have been tested on the 16-object version of the dataset (MIT-16) [60, 133, 161], while others have used the extended 20-objects version (MIT-20) [90, 136, 133, 20, 28]. Shen and Yeo, in [138], only provided results for 13 objects of the MIT dataset.

Barron and Malik presented an extension of the MIT dataset in [20]. In this extended dataset, photometric stereo was used to estimate the depth and the spherical harmonic illumination of each of the objects in the original MIT dataset [71]. The authors further extended the dataset in [19] by applying to the objects different illumination models from the sIBL Archive<sup>‡</sup>. An example of the extended dataset is shown in Figure 2.6.

### Synthetic datasets

Building synthetic datasets considerably simplifies the process of acquiring ground truth data for different intrinsic scene components such as reflectance, shading, shape or the color and position of the illuminant. Although rendering engines use approximate reflection models, they are able to reproduce complex scenes which are difficult to differentiate from actual real scenes.

Bousseau *et al.*, in [32], provided the first synthetically rendered scene for intrinsic image evaluation (see Figure 2.7). This image was also used in posterior works [136, 138].

The MPI-Sintel dataset [35] is a new optical flow dataset where 35 sequences displaying different environments, characters/objects, and actions were extracted from a 3D animated short film. It was not specifically designed for the purpose of intrinsic

<sup>‡</sup><http://www.hdrlabs.com/sibl/archive.html>

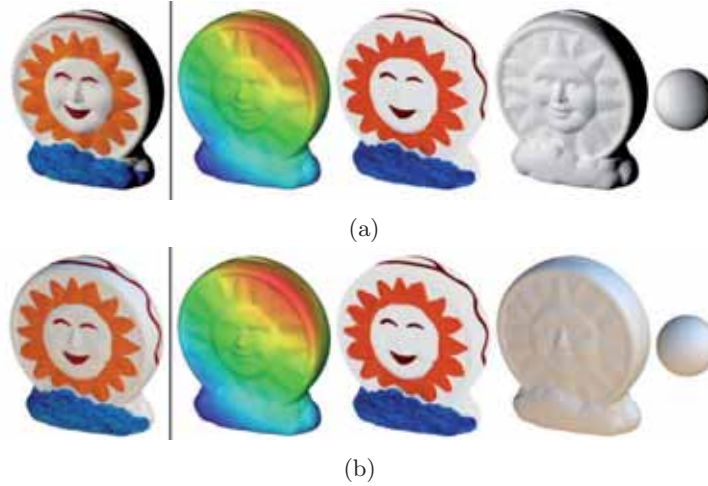


Figure 2.6: Barron and Malik extended the MIT dataset [71]. Example of extensions provided for the sun object. (a) Estimating shape and illuminant information [20]. (b) Adding complex illumination models [19].

image evaluation, but it provides multiple useful groundtruth information which has already been used for quantitative evaluation in some works [38, 159]. Figure 2.8 shows an example of the ground truth information that can be found in this dataset.

### Crowdsourced datasets

Bell *et al.*, in [28], built a large dataset for intrinsic image decomposition using internet real-world photos of indoor scenes and crowdsourced annotations about the relative reflectance of pairs of pixels in each image. Their dataset contains over 5000 images with an average of 100 human judgements per image, and a set of about 400 densely annotated images (around 900 human annotations per image). The authors argued that humans are not good at making absolute judgements. Accordingly, they asked the observers which of the two points had a darker surface color and allowed three possible answers (“the first”, “the second” and “both”). Subjects were also asked about the confidence of their answers, and this information was used to define weights for each judgement. Noisy images were removed, as well as greyscale photographs and images with exaggerated camera effects. The skills of the subjects who performed the annotations were also tested using “sentinel” objects for which the real answer was known, allowing the authors to discard the answers of annotators who misunderstood the task or did not do it correctly.

The main advantage of this dataset is that it contains a large number of natural images, making it possible to evaluate the methods for a big range of different scenes. However, this dataset does not include ground truth intrinsic images such as they were defined in [23], but just a set of sparse annotations about the reflectance of the images based on imperfect human judgement.

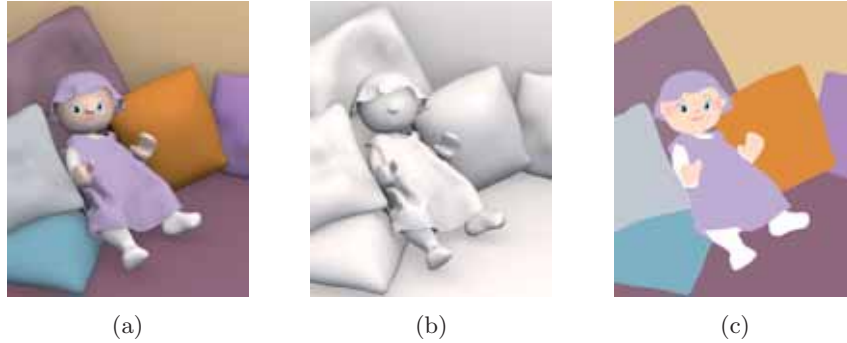


Figure 2.7: Synthetic image of a doll in [32]. (a) Original image. (b) Shading image. (c) Reflectance image.

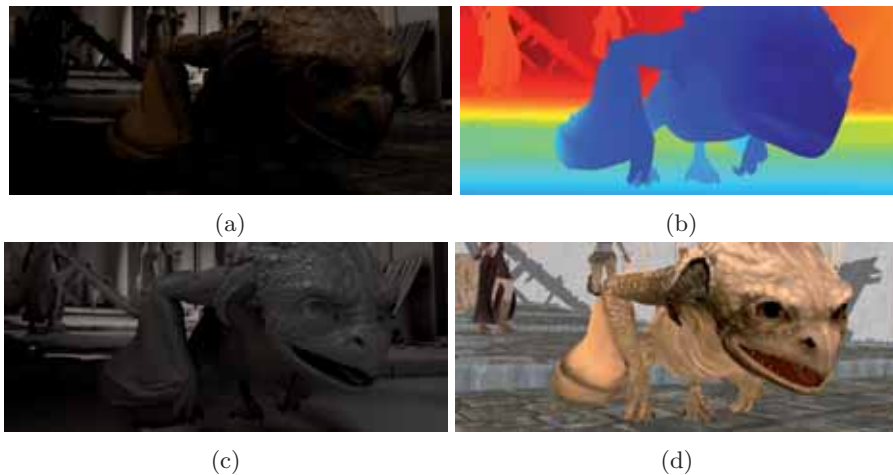


Figure 2.8: Example of MPI-Sintel dataset groundtruth [35]. (a) Original image. (b) Normal map. (c) Shading image. (d) Reflectance image.

### 2.4.3 Metrics

Even when ground truth information is available, methods on intrinsic image decomposition need a standardized way to measure how close their intrinsic image estimates are with respect to the ground truth data. Therefore, it is necessary to define similarity measures in order to quantitatively evaluate and compare the different intrinsic image approaches.

The first quantitative evaluation of intrinsic images can be found in [145], where Tappen *et al.* used the mean squared error (MSE) between the estimated and the ground-truth shading images,

$$\text{MSE}(I_1, I_2) = \frac{1}{|I_1|} \operatorname{argmin}_{\alpha} \|I_1 - \alpha I_2\|_2^2, \quad (2.4)$$

where  $I_1$  and  $I_2$  are the images to be compared,  $\alpha$  provides intensity-invariance to the

metric and  $|I_1|$  is the number of pixels in the image  $I_1$  and is used as a normalization value (the size of both images is assumed to be the same). This error measure has also been used in further works [90, 133, 38, 20, 19, 21]. The main problem of MSE for the evaluation of edge classification approaches is that it is a global metric (*i.e.* takes into account the whole image). In this same direction, Grosse *et al.* also suggested in [71] that the MSE metric is too strict and that misclassifying a single edge in a little part of the image can lead to big MSE scores. For instance, in Figure 2.9 we have represented a plausible scenario where we are given an image containing a single color and an intensity gradient caused by illumination. We use a Retinex-like method which wrongly classifies an intensity edge as a reflectance change in the center of the image. Although the method has just misclassified a line of pixels, the MSE metric considers that at least half of the pixels in the computed image are wrong, because it computes the error in the whole image.

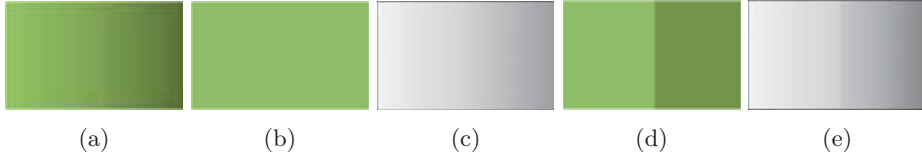


Figure 2.9: Example to illustrate the behaviour of different error metrics, given an original image (a), its ground truth reflectance (b) and shading (c) and the computed estimates for reflectance (d) and shading (e) using a Retinex-like method which misclassifies an image derivative.

Grosse *et al.*, in [71], proposed a new metric for intrinsic image evaluation called the local mean square error (LMSE). This error measure is defined as the average of the MSE calculated on overlapping image patches,

$$\text{LMSE}(I_1, I_2) = \frac{\sum_w \operatorname{argmin}_\alpha \|I_1^w - \alpha I_2^w\|_2^2}{\sum_w \|I_1^w\|_2^2}, \quad (2.5)$$

where  $w$  represents an image window.

The problem of the misclassified edge resulting in a big MSE score illustrated in Figure 2.9 is avoided with the LMSE metric, since only the windows including a part of the misclassified edge accumulate some error, which is later averaged by the total number of windows. Notice that any other window which does not include the wrong reflectance edge will perfectly match its GT counterpart, since both windows only differ by an intensity scalar value and the MSE metric is scale-invariant.

The LMSE metric has been widely used for method evaluation in the intrinsic image literature [90, 136, 138, 60, 20, 19, 133, 59, 102, 21, 38]. However, its suitability for intrinsic image evaluation has also been discussed by some authors [90, 60, 133].

In [90], Jiang *et al.* argued that the LMSE is biased towards the mean values of the compared images and proposed the absolute local mean square error (aLMSE), where the mean values of both images in the patch are subtracted to supply the LMSE with insensitivity to variations in mean intensity,

$$\text{aLMSE}(I_1, I_2) = \frac{\sum_w \operatorname{argmin}_\alpha \|(I_1^w - \mu_1^w) - \alpha(I_2^w - \mu_2^w)\|_2^2}{\sum_w \|(I_1^w - \mu_1^w)\|_2^2}. \quad (2.6)$$

The authors also proposed a correlation-based metric to measure similarity. This metric can either be applied globally to the images,

$$\operatorname{Corr}(I_1, I_2) = \frac{\mathbb{E}[(I_1 - \mu_1)(I_2 - \mu_2)]}{\sigma_1 \sigma_2}, \quad (2.7)$$

or calculated in local windows of the image and averaged,

$$\operatorname{LCorr}(I_1, I_2) = \frac{1}{|W|} \sum_{w \in W} \frac{\mathbb{E}[(I_1^w - \mu_1^w)(I_2^w - \mu_2^w)]}{\sigma_1^w \sigma_2^w}, \quad (2.8)$$

where  $|W|$  is the number of windows. Correlation and the aLMSE metric are also used in other works [136, 133]. Serra *et al.*, in [133], suggested that the LMSE metric benefits Retinex-like approaches but may prejudice other energy optimization methods which do not classify image derivatives.

Gehler *et al.*, in [60], claimed that the LMSE is not robust to outliers, and proposed to use the average rank of the algorithm to evaluate the methods. This measure consists in ranking the results of the different methods for each of the images in the dataset, and then averaging the rankings of each method.

In a different direction, Chen and Koltun, in [38], defined the structural dissimilarity index (DSSIM), which is a variant of the perceptually based structural similarity index (SSIM) proposed in [156],

$$\operatorname{SSIM}(I_1, I_2) = \frac{(2\mu_1\mu_2 + C_a)(2\sigma_{12} + C_b)}{(\mu_1^2 + \mu_2^2 + C_a)(\sigma_1^2 + \sigma_2^2 + C_b)}, \quad (2.9)$$

where  $\mu_i$  is the mean intensity of image  $I_i$  and  $\sigma_i^2$  is its variance. The term  $\sigma_{ij}$  indicates the covariance between images  $I_i$  and  $I_j$ , and the scalar values  $C_a$  and  $C_b$  provide stability to the system. If these values were set to 0, the SSIM would produce unstable results when either  $\mu_1^2 + \mu_2^2$  or  $\sigma_1^2 + \sigma_2^2$  were very close to zero. The main idea underlying this metric is that human beings are good at isolating features that provide structural information of the scene. Given two images, the SSIM combines different comparisons about their luminance, contrast and structure.

The Mean SSIM is defined as an averaged sum of SSIM applied to local windows on the images,

$$\operatorname{MSSIM}(I_1, I_2) = \frac{1}{|W|} \sum_{w \in W} \operatorname{SSIM}(I_1^w, I_2^w), \quad (2.10)$$

and the DSSIM is defined to transform this similarity measure into a dissimilarity measure, in the same line as for the MSE or the LMSE,

$$\operatorname{DSSIM}(I_1, I_2) = \frac{1 - \operatorname{MSSIM}(I_1, I_2)}{2}. \quad (2.11)$$

However, it has already been questioned whether the SSIM is better at reproducing human perception than the MSE. In fact, Dosselmann and Yang [43] demonstrated that SSIM is closely related to the MSE.

Another metric based on human perceptual judgement, the weighted human disagreement rate (WHDR), was introduced recently in [28]. Bell *et al.* defined the WHDR as

$$\text{WHDR}_\delta(J, R) = \frac{\sum_{i=1}^{|J|} w_i \cdot (J_i \neq \hat{J}_{i,\delta}(R))}{\sum_{i=1}^{|J|} w_i}, \quad (2.12)$$

where  $R$  is the output reflectance and  $w_i$  and  $J_i$  are the confidence weights and given judgements, respectively, obtained from human observations. Human judgements  $J_i$  can be equal to 1 when the darkest point has a lighter surface reflectance, equal to 2 when the darkest point has a darker surface reflectance, or equal to E when both points have equal reflectance intensities.  $\hat{J}_{i,\delta}$  is the judgement predicted by the algorithm being evaluated. In order to transform algorithm reflectances into judgements, the authors threshold differences between the points used in the human judgement in  $R$ :

$$\hat{J}_{i,\delta}(R) = \begin{cases} 1 & : \frac{R_{2,i}}{R_{1,i}} > 1 + \delta \\ 2 & : \frac{R_{1,i}}{R_{2,i}} > 1 + \delta \\ E & : \text{otherwise} \end{cases} \quad (2.13)$$

where  $R_{j,i}$  is the reflectance sampled at point  $j$  for judgement  $i$ .

The main drawback of this measure is that it only evaluates an image using a sparse set of pairs of pixels for which human judgements exist. However, it is not clear if this sparse set of pixels is representative of the performance of the algorithm on the whole image. Moreover, the WHDR only evaluates relative information about the reflectance of these pairs of pixels, but does not evaluate the global coherence of the image. This means that we could always find a reflectance estimate where the reflectance values are not coherent with the actual scene, but the selected reflectance ratios coincide with those of the human judgements.

In order to evaluate and compare intrinsic components other than shading and reflectance, other similarity measures have been used. For instance, Barron and Malik, in [20, 19], evaluated the estimated shape output of their method with the mean absolute error (MAE),

$$\text{MAE}(I_1, I_2) = \frac{1}{|I_1|} \operatorname{argmin}_\alpha \|I_1 - (I_2 + \alpha)\|_1, \quad (2.14)$$

where  $\alpha$  provides shift-invariance. The MAE metric is a global metric very similar to the MSE metric. Therefore, the advantages and disadvantages of both essentially coincide.

#### 2.4.4 Discussion

The evaluation of intrinsic images has proved challenging so far. In the field of intrinsic images, we need standard similarity measures which overcome the problems of the existing metrics. Moreover, large datasets containing complex scenes and ground truth data for multiple intrinsic images are necessary. In this direction, a crowdsourced dataset [28] has been proposed recently.

Additionally, in Chapter 5 we present two new datasets. One of them includes ground truth information about the illuminant of the scene and the camera sensors. These two factors influence the image values during the acquisition process, but have usually been ignored in the field of intrinsic image decomposition. The other dataset [24] uses synthetic data in order to build a dataset containing complex scenes under multiple illumination conditions.

Table 2.4 summarizes how the existing methods in intrinsic image decomposition have used the evaluation mechanisms described in this section. Finally, Table 2.5 summarizes different aspects of the existing datasets such as the nature of their data or the type of ground truth information they provide.

## 2.5 Conclusion and Perspectives

Intrinsic image methods have originally focused on the decomposition of a single image into its reflectance and shading components, putting emphasis on the assumptions they make about regularities of the scenes and the image formation process in order to constrain the problem. Recently, the improvement of optimization methods and the development of technology have made intrinsic image decomposition evolve towards the inclusion of extra input information and the joint estimation of other intrinsic features of the scene. Therefore, in the last few years, the gap between the original intrinsic image decomposition problem and other problems in computer vision which also use intrinsic information such as color constancy [63], shape from shading [45], highlight removal [13], etc. has been notably reduced.

The emergence of new methods that include extra input information and jointly estimate diverse intrinsic components makes it necessary to build larger datasets including ground truth data of many different intrinsic components of a scene. However, as we have seen in Section 2.4, it is a tough problem given the actual complexity of building datasets for intrinsic image evaluation consisting of natural scenes. Crowdsourced datasets could be an alternative for building huge datasets of natural images. Actually, the only dataset of this kind [28] only checks a sparse set of pixels of the computed reflectance images using human perception. It is still unclear if such a dataset can be adjusted to accurately evaluate multiple intrinsic factors. In a different direction, synthetic image datasets [35] provide realistic representations of the real world and have been successfully used in other problems [116, 150]. In the computer graphics software used to render synthetic images, ground truth information can be easily isolated and modified, allowing the creation of multiple different images from a single scene. Accordingly, synthetic datasets may play a key role in the future of intrinsic image evaluation.

Another difficulty in the field of intrinsic images is the definition of an accurate similarity measure. Although many different metrics exist in the literature of intrinsic image decomposition, as we have seen in Section 2.4, all of them present multiple disadvantages. The most used similarity measure so far has been the LMSE metric [71], but it has also been hardly questioned [90, 60, 133]. Recent similarity measures [38, 28] are based on human perception. However, one of these metrics has been also questioned [43] for its inaccurate representation of human perception, and the other

one uses a big amount of human observations, which makes it impractical. Finding a standard similarity measure to evaluate intrinsic images is one of the most important needs in this field.

The future of intrinsic image decomposition includes the joint estimation of multiple intrinsic components (such as the shape, the color and direction of the illuminant [21] or the response of the camera sensors [134]) and other useful information (like semantic segmentation [154], optical flow [99] or scene structure from multi-view stereo [102]). Although adding new terms to the basic formulation of the problem may seem to add complexity to the problem, these terms are in fact highly dependent on each other. These dependencies and the different assumptions we make on each intrinsic component help us constrain the space of possible solutions. Furthermore, the use of advanced inference techniques such as high-order markov random fields [154] and deep learning methods [144] have proved to be useful in order to successfully estimate multiple intrinsic components.



Work	Quantitative Results	Datasets	Metrics
Funt'92 [57]			
Sinha'93 [141]			
Bell'01 [27]		(1)	
Weiss'01 [157]		(2)	
Finalyson'04 [49]			
Matsushita'04 [119]			
Olmos'04 [122]			
Tan'04 [143]			
Tappen'05 [146]			
Tappen'06 [145]	✓	(3)	(a)
Bousseau'09 [32]		(10)	
Jiang'10 [90]	✓	(4,6)	(b,c,d,e)
JShen'11 [136]	✓	(6)	(a,b,c)
LShen'11 [138]	✓	(5)*	(b)
Gehler'11 [60]	✓	(5)	(b,f)
Barron'12a [20]	✓	(7)	(a,b,h)
Barron'12b [19]	✓	(7)	(a,b,h)
Garces'12 [59]		(6,10)	
Lee'12 [105]			
Serra'12 [133]	✓	(5,6)	(a,b,c,d)
Laffont'12 [102]	✓	(10)	(b)
Tang'12 [144]		(2)	
Zhao'12 [137, 161]		(10)	
Barron'13 [21]	✓	(7,8)	(a,b,h)
Chen'13 [38]	✓	(8,9)	(a,b,g)
Vineet'13 [154]		(8)	
Chang'14 [37]	✓	(5,6)	(a,b)
Jeon'14 [89]		(8)	
Ye'14 [159]	✓	(9)	(b)
Kong'14 [99]	✓		(b)
Bell'14 [28]	✓	(6,11)	(b,i,f)

Table 2.4: Classification of the methods based on the evaluation tools that the authors have used. Databases: (1) Psychophysics [55]. (2) Yale Face [61]. (3) Crayola [145]. (4) BOLD [90]. (5) MIT - 16 objects [71]. (6) MIT - 20 objects [71]. (7) extended MIT [20]. (8) NYU Depth [139]. (9) MPI-Sintel [35]. (10) images from Flickr. (11) Intrinsic Images in the Wild [28]. Metrics: (a) MSE (Eq. 2.4). (b) LMSE (Eq. 2.5). (c) aLMSE (Eq. 2.6). (d) Corr (Eq. 2.7). (e) LCorr (Eq. 2.8). (f) Average Rank. (g) DSSIM (Eq. 2.11). (h) MAE (Eq. 2.14). (i) WHDR (Eq. 2.12). \* Only results for 13 objects are provided.

Dataset	Acquisition		Ground Truth				
	Data	Objects	Shad. & Refl.	Specularities	Depth	Illuminant	Sensors
Crayola [145]	Laboratory	46	✓				
MIT [71]	Laboratory	20	✓	✓			
extended MIT [20]	Synthetic	20	✓	✓	✓	✓	
MPI-Sintel [35]	Synthetic	35*	✓		✓		
Bell <i>et al.</i> [28]	Natural	5000	✓				

Table 2.5: Comparison of the existing datasets for intrinsic image evaluation. \* Sequences with different numbers of frames (50 on average).

# Chapter 3

## Shades and Names of Color for Intrinsic Image Estimation

In this chapter we propose a method for intrinsic image estimation that aims to overcome the shortcomings of pure edge-based methods by introducing strong surface descriptors such as a color-name descriptor, which introduces high-level considerations resembling top-down intervention. We also use a second surface descriptor, termed color-shade, which allows us to include physical considerations derived from a model of image formation that captures gradual color surface variations. Both color cues are combined by means of a Markov Random Field. Figure 3.1 shows how our method performs with a natural image which has been previously used in other works [32, 138].

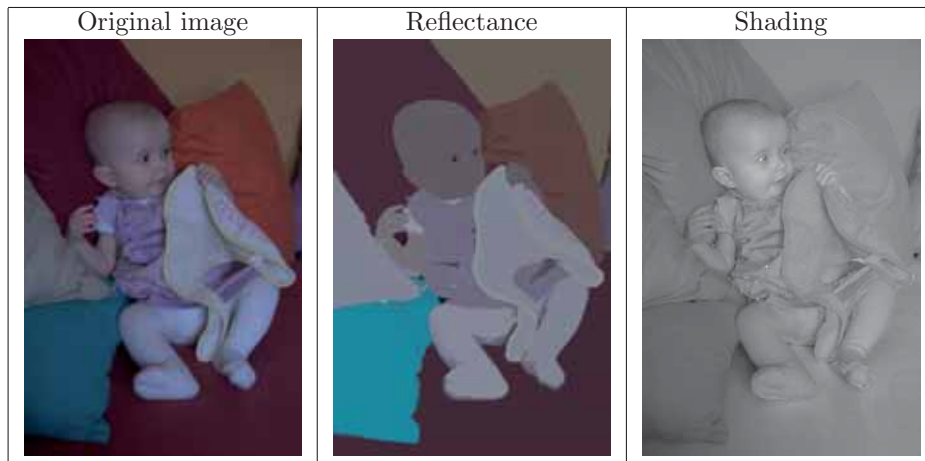


Figure 3.1: Shading and reflectance images recovered with our method.

We focus on the decomposition of an image into its reflectance and shading components. As we have previously seen in Chapter 1, recovering these two characteristics from a single image  $I$  amounts to estimate a reflectance image,  $I_{RefI}$ , and a shading

image,  $I_{Shad}$ , such that

$$I(x, y) = I_{Shad}(x, y) \cdot I_{Ref}(x, y). \quad (3.1)$$

As it has already explained in the previous chapters, this problem is clearly ill-posed because the number of unknowns is higher than the number of equations.

This chapter is organized as follows. The motivation underlying the proposed method of intrinsic image decomposition is introduced in Section 3.1. Section 3.2 describes the color cues used to recover reflectance and shading images as explained in Sections 3.3 and 3.4. Finally, concluding remarks are given in Section 3.6.

### 3.1 Motivation

From the analysis, in Chapter 2, of previous approaches on intrinsic image decomposition we draw the following conclusions:

- The exclusive use of edges to recover surface reflectance is not sufficient since a small missclassified edge could provoke an error over a wide area. Therefore, we argue that surface attributes such as color and texture could be essential cues to improve edge-based proposals.
- Results from methods that include user interaction suggest that top-down intervention yields a clear advantage for dealing with the ill-posed nature of reflectance recovery. Hence, we argue for the need of high-level attributes to describe image content.
- Moreover, few efforts have been made to exploit the information derived from the assumption that image formation obeys a specific physical model. Intrinsic image algorithms could benefit from these models since they account for changes in image appearance due to the geometry and illumination of the scene.

In view of these considerations, we propose the introduction of color surface attributes based on color names instead of an edge-based analysis. These attributes provide high-level information resembling top-down intervention in the reflectance recovery. Afterwards, we add a second descriptor, termed color-shade, that allows us to take into account physical considerations on color surface variations due to the geometry and lighting of a scene. This descriptor, which assumes Shafer's dichromatic reflection model for image formation [135], is introduced to address the lack of stability of the color-name descriptor in the presence of strong variations in the geometry or illumination of a scene. Color-name and color-shade descriptions are finally combined by means of a Markov Random Field.

### 3.2 Our approach

In this section we first introduce the color-name and color-shade descriptors. We then outline the conditional inference approach that we adopt to combine these color cues. This conditional inference will be further detailed in Section 3.3.

### 3.2.1 Color-name descriptor

Color-name descriptors associate the linguistic terms humans use for describing objects to colors in an image. Basic color names were first defined by Berlin and Kay [31]. They were deduced from a large anthropological study based on speakers of 20 different languages and specific documentation from 78 other languages. The authors concluded that the universal basic color terms defined in most evolved cultures\* are 11. Subsequent psychophysical experiments have generated data that allow these basic names to be accurately specified [29] and computationally implemented [30]. Accordingly, color-naming models provide perceptually-based quantizations of the RGB color space, which present a higher discrimination power with respect to usual chromaticities, as proven for classification tasks [70].

We use the color-naming model proposed by Benavente *et al.* [30], where a color name category is modeled as a fuzzy set with a membership function that, given a color sample, assigns a value between 0 and 1 to represent the color-name membership. The model uses the set of names proposed by Berlin and Kay [31], namely  $\mathcal{N} = \{black, white, red, green, yellow, blue, brown, purple, orange, pink, grey\}$ . Since the model forces the sum of all memberships to be 1 for any pixel, the membership values can be considered as probabilities. The color-name descriptor of a pixel  $\mathbf{p}_i$ , denoted  $ND(\mathbf{p}_i)$ , is a 11-dimensional real-valued vector whose components are the probabilities of labeling the given pixel with each one of the color names in  $\mathcal{N}$ . More explicitly,

$$ND(\mathbf{p}_i)_j = Prob(N_j|\mathbf{p}_i), \forall j = 1, \dots, 11, \quad (3.2)$$

where  $N_j$  is the  $j$ -th color name in the set of basic color names  $\mathcal{N}$ . Figure 3.2(a) shows the volumes of the RGB space where each of the 11 color names have probability 1.

The color-name descriptor has two interesting properties. First, it is relatively invariant to small photometric changes since wide areas of a single reflectance surface assume the same label and small changes in shading only cause gradual changes in the descriptor. Next, it provides a sparse representation of color since very few coordinates of the 11-dimensional vectors are non-zero at the same time (usually up to three). Since this descriptor yields the labels of a conditional inference labeling problem (see Section 3.3), we only allow three coordinates to be non-zero and we further discretize the probability vector by quantizing the coordinates to  $\{0, 0.25, 0.5, 0.75, 1\}$  while keeping the constraint that they sum to 1. This means that a color can be described with a maximum of three names (for instance, greyish blue-green), which is a perceptually consistent constraint since in the model very few colors are in the boundary of four color names [30]. With such restrictions a total of 671 labels are theoretically possible (see Appendix A). However, the majority of these labels are never found in practice because in the color space no color border is shared by more than 3 or 4 colors. Therefore, many labels defining unfeasible combinations of colors, such as bluish yellow-purple for example, are never used. In the end, only 250 different labels are actually used. Thus, just considering labels with up to three positive coordinates is enough to accurately describe the whole RGB space. In the next sections we prove that this set of labels is a reliable sparse representation of color to

---

\*The authors state that “there appears to be a positive correlation between general cultural complexity (and/or level of technological development) and complexity of color vocabulary.”

recover reflectance. In Figure 3.2(b) we show an example of the color-name labels assigned to an image.

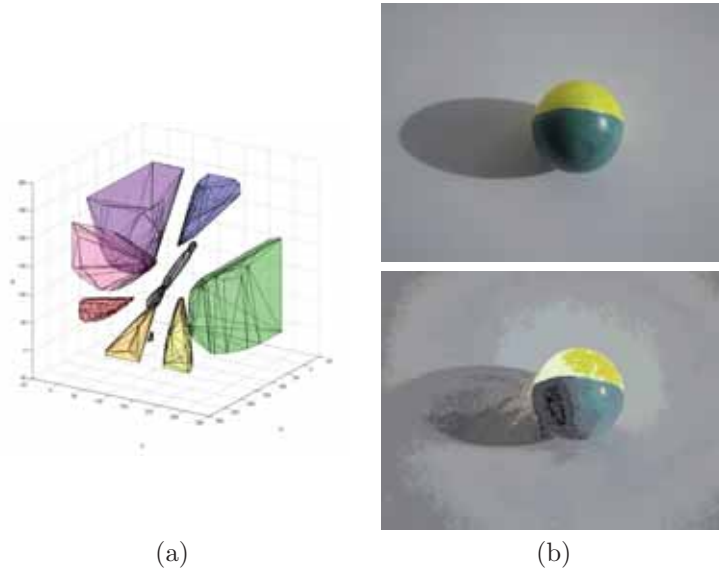


Figure 3.2: Color-name descriptor. (a) Plotted volumes represent those values in the RGB color space with probability equal to 1 of being one of the 11 universal colors according to [30]. The space between the volumes corresponds to vectors for which at least two coordinates are positive. (b) Image labeled with the color-name descriptor.

### 3.2.2 Color-shade descriptor

The color-shade descriptor is based on the method of Vazquez *et al.* [151], where the authors propose to describe scene reflectances using a ridge analysis of the color distributions (RAD). Vazquez *et al.* define a ridge as a set of points connecting the local maxima of a color distribution in the RGB histogram space. In Figure 3.3(b), we show a 3D representation of the color distribution of the image in 3.3(a). The corresponding ridges (connected maxima) detected by the RAD method on the 4D color histogram distribution are shown in Figure 3.3(c), where we can see four ridges corresponding to the blue, red, orange and white parts of the image. By looking at the ridges, we can see how smooth variations of shade are represented for each color. For example, the white ridge spans colors from the lightest white to the darker gray present in the shadowed part of the object. In the ideal case, the RAD method provides a single ridge for each reflectance surface of the image.

The physical model underlying the RAD method is the dichromatic reflection model described by Shafer in [135]. In this model, all the color variations of a surface, including shading effects and highlights, span a 2D plane in the RGB space which is defined by two vectors: one in the direction of the surface's albedo, and the other in the direction of the illuminant. Hence, the dichromatic model, and, therefore, the

RAD method, provide a compact representation of all the variation that a single-color surface can present due to illumination changes and the geometry of the scene.

Given an image, the RAD method returns a set of ridges  $\mathcal{R} = \{R_1, \dots, R_n\}$ . From this set of ridges, we define the color-shade descriptor of an image pixel  $\mathbf{p}_i$  as

$$SD(\mathbf{p}_i) = \operatorname{argmin}_{R_j \in \mathcal{R}} (\operatorname{dist}(\mathbf{p}_i, R_j)), \quad (3.3)$$

where  $\operatorname{dist}(\mathbf{p}_i, R_j)$  represents the Euclidean distance between the RGB value  $\mathbf{p}_i$  and the nearest point of the ridge  $R_j$ .

Thus, ridges provide useful information for enhancing the color-name representation and allow us to deal with the variations of color names in the presence of strong illumination effects, *i.e.* shading and highlights. For example, two pixels belonging to the same reflectance object but with very different RGB values, *e.g.* one in a shadowed part of the object, the other in a brighter part, are connected by their nearest ridge. Following this approach, one can consistently name pixels within a single reflectance area allowing for shading changes.

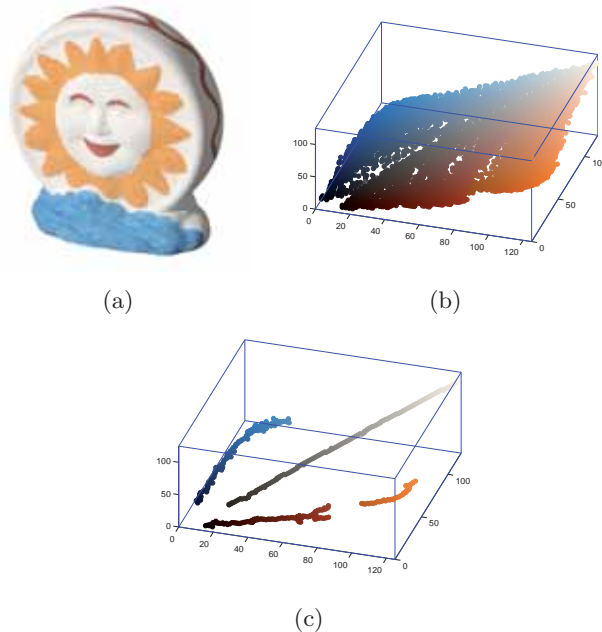


Figure 3.3: Color-shade descriptor. An image from the MIT dataset (a), its color distribution (b), and the ridges detected by the RAD method (c).

### 3.2.3 Method outline

Our algorithm is based on the assumption that the reflectance of a single material can be described by a unique color name provided by the descriptor introduced in Section 3.2.1. Spatial coherence for this descriptor is then achieved by propagating evidence

through a MRF. The color-name descriptor provides an accurate color edge localization, where relevant color edges are perfectly located by changes in color names, as well as a meaningful surface interpretation based on standard prior knowledge compiled from psychophysical data.

However, when strong shading variations occur, irrelevant edges can appear within a surface with uniform reflectance, causing an undesirable over-segmentation. To deal with this problem, we break the homogeneity of our MRF in order to incorporate the physical information that the color-shade descriptor provides. This step stems from the assumption that changes in shading within an area of uniform reflectance yield to connected distributions of points in the RGB histogram and prevents efficiently the excessive segmentation of color names in shaded and near highlight areas.

In a final stage, once information from names and shades has been propagated, we modify the reflectance description provided by the MRF to match the intensities of the recovered reflectance to those of the original image. A scheme of the method is given in Figure 3.4.

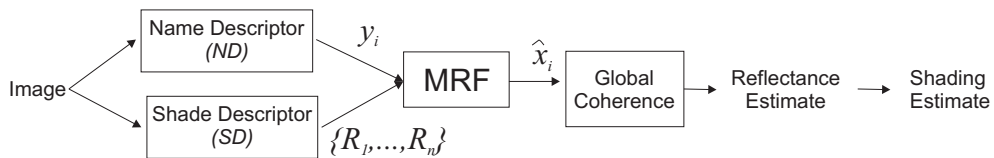


Figure 3.4: Block diagram of our method for intrinsic image estimation.

### 3.3 Reflectance recovery using MRF inference

In this section we present how our method uses MRF inference to estimate an intrinsic reflectance image combining the two color cues presented in the previous section.

Let  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  be the graph that represents the input color image, where the set of vertices  $\mathcal{V}$  correspond to random variables  $x_i$  associated to the set of pixels of the image (one node for each pixel), and  $\mathcal{E}$  is the set of undirected edges representing relationships between pairs of adjacent pixels (using a 4-neighborhood system). The set of maximal cliques  $\mathcal{C}l$  is formed by the edges of the graph  $\{x_i, x_j\}$ , where  $i$  and  $j$  are adjacent pixels, and the cliques of the form  $\{x_i, y_i\}$ , for each pixel  $i \in \mathcal{V}$ , where  $y_i$  stands for the observation at pixel  $i$ .

Both random variables  $x_i$  and observations  $y_i$  are reflectance values as expressed by the color-name descriptor outlined in the previous section. Accordingly, the set of labels  $\mathcal{L}$  is a set of 11-dimensional vectors.

The energy function of our MRF is

$$E(\mathbf{x}) = \mu \sum_{i \in \mathcal{V}} D(x_i, y_i) + (1 - \mu) \sum_{\{i, j\} \in \mathcal{E}} V(x_i, x_j), \quad (3.4)$$

where  $D(x_i, y_i)$  is the singleton potential defined on each pixel,  $x_i$ , and  $V(x_i, x_j)$  is the pairwise potential defined on a pair of neighboring pixels. The contribution of both terms in the global energy is weighted using a parameter  $\mu \in [0, 1]$ .



The  $\alpha$ -expansion graph cut algorithm [14] presented in [33] is used to find the labeling  $\hat{\mathbf{x}}$  that minimizes the energy term expressed by Equation 3.4.

### 3.3.1 Singleton potential: color name

The singleton potential  $D(x_i, y_i)$  measures to which extent the labeling  $\mathbf{x}$  fits the observed data  $\{y_i\}_{i \in \mathcal{V}}$ . In practice, this potential can be interpreted as the cost of assigning  $x_i$  a label different from the label of observation  $y_i$ .

For computing the singleton potential, we chose the  $L_1$  distance in the Euclidean space of 11-dimensional probability vectors:

$$D(x_i, y_i) = \|x_i - y_i\|_1, \forall i \in \mathcal{V}. \quad (3.5)$$

We also tried to use other distances, but they did not lead to significant improvements.

### 3.3.2 Pairwise potential: color shade

In classic MRFs, the pairwise potentials  $V(x_i, x_j)$  measure the smoothness of the labeling  $\mathbf{x}$  and can be interpreted as the cost of assigning different labels to neighboring pixels. Our first idea is to define these potentials using the Euclidean distance,

$$V(x_i, x_j) = \|x_i - x_j\|_1, \forall (i, j) \in \mathcal{E}. \quad (3.6)$$

However, in the pairwise potential of our MRF, we also include information from the color-shade descriptor by weighting the value of the distance between each pair of neighboring pixels. The main idea underlying this formulation is that pairs of pixels belonging to the same ridge should belong to the same surface and therefore should share similar labels: the cost of holding different names should be higher for neighboring pairs of pixels whose observed RGB values belong to the same ridge. Following this idea, we define the pairwise potential as

$$V(x_i, x_j) = \omega_{ij} \|x_i - x_j\|_1, \quad (3.7)$$

where  $(x_i, x_j)$  are the labels of a pair of neighboring pixels and  $\omega_{ij}$  weights the classical smoothness term according to the relative position of the RGB values  $\mathbf{p}_i$  and  $\mathbf{p}_j$  of pixels  $i, j$  and the ridges of the color-shade descriptor as explained below.

Let  $\pi(\mathbf{p}_i)$  be the orthogonal projection of the pixel value  $\mathbf{p}_i$  on its associated ridge  $SD(\mathbf{p}_i)$  and let  $\theta_{ij}$  be the angle formed by the lines  $(\overline{\mathbf{p}_i \mathbf{p}_j})$  and  $(\overline{\pi(\mathbf{p}_i) \pi(\mathbf{p}_j)})$ . Given a pair of pixel values  $(\mathbf{p}_i, \mathbf{p}_j)$  and the set of ridges of the image, we distinguish three cases of relative position between these pixels (Figure 3.5 illustrates these cases).

*Case A:*  $\omega_{ij} = \alpha$  if the two pixels lie on two different ridges, *i.e.*  $SD(\mathbf{p}_i) \neq SD(\mathbf{p}_j)$ ;

*Case B:*  $\omega_{ij} = \beta$  if the two pixels lie on the same ridges, *i.e.*  $SD(\mathbf{p}_i) = SD(\mathbf{p}_j)$ , but the direction they determine is not parallel to the ridge, *i.e.*  $\theta_{ij} > thr$ , where  $thr$  is a parameter fixed once for all;

*Case C:*  $\omega_{ij} = \gamma$  if the two pixels lie on the same ridges, *i.e.*  $SD(\mathbf{p}_i) = SD(\mathbf{p}_j)$ , and the direction they determine is parallel to the ridge, *i.e.*  $\theta_{ij} \leq thr$ .

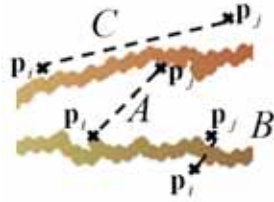


Figure 3.5: Schema of the three different scenarios we considered in the description of the pairwise potential.

This weight assignation is summarized below,

$$\omega_{ij} = \begin{cases} \alpha & : SD(\mathbf{p}_i) \neq SD(\mathbf{p}_j). \\ \beta & : SD(\mathbf{p}_i) = SD(\mathbf{p}_j), \\ & \theta_{ij} > thr. \\ \gamma & : SD(\mathbf{p}_i) = SD(\mathbf{p}_j), \\ & \theta_{ij} \leq thr. \end{cases}$$

The choice of the parameters  $\alpha$ ,  $\beta$ , and  $\gamma$  must be consistent with the idea that the cost of holding different names should be higher for pairs of pixels whose observed RGB values belong to the same ridge. In particular, this cost should dramatically increase if, in addition, they form a segment whose direction is collinear to the ridge's direction because this corresponds to the paradigmatic case of two pixels belonging to the same reflectance object but with different shadings. Accordingly,  $(\alpha, \beta, \gamma)$  should verify the inequalities  $\alpha \leq \beta \leq \gamma$  and  $\alpha \ll \gamma$ . In Figure 3.6 we illustrate how the different cases affect the MRF in the model.

### 3.3.3 MRF output

The output of the MRF consists of an array of probability vectors. However, what we expect to recover are reflectance values (*i.e.* RGB triplets). Accordingly, we need a way to set a link between RGB and probability values.

Since we first discretize the probability vectors, many RGB values are mapped to a single vector by the color-name descriptor. This provides a partition  $\coprod_{v \in \mathcal{L}} S_v$  of the RGB cube, where  $S_v$  is the convex set of RGB values associated to label  $v$ . The inverse mapping is defined by associating each probability vector (*i.e.* label  $v \in \mathcal{L}$ ) with the center of mass of the convex region it defines in the RGB space.

## 3.4 Adding global scene coherence

Up to this point we have used local and semi-local color information to recover reflectance estimates. Our MRF outputs a representative RGB value for each area of uniform reflectance. However, the output image lacks global consistency. In particular, the intensity of one region in our estimated reflectance image may not be coherent

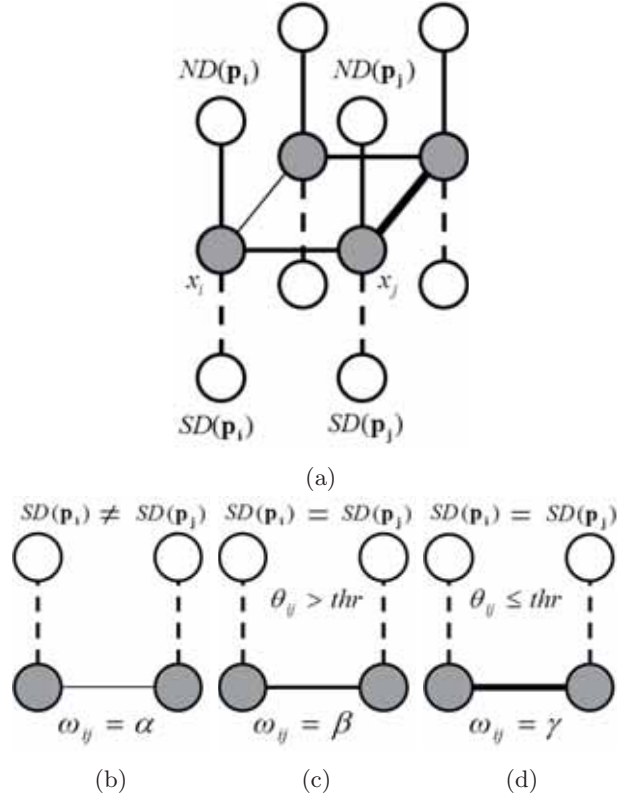


Figure 3.6: Schema of our Markov random field. (a) Each pixel  $x_i$  has two observations, one from the color-name descriptor ( $ND$ ) and another from the color-shade descriptor ( $SD$ ). The continuity between two pixels is enforced in the pair-wise potential. The width of the lines represents the weight we assign to each of the edges in order to penalize label discontinuities. These weights are locally modified according to the different cases defined by the color-shade observations: (b) the pixels belong to different ridges (Case A). We assign a small weight to this edge, since no continuity is expected between these two pixels in the reflectance image. (c) The pixels belong to the same ridge but the line they form is not parallel to the ridge (Case B). We assign a medium weight to this edge. We want to penalize different neighboring labels moderately as a consequence of the unclear conclusion we draw from the color-shade descriptor, allowing certain flexibility for label changes in this scenario. (d) The pixels belong to the same ridge and the line they form is parallel to the ridge (Case C). We assign a big weight to this edge, since we do not want these pixels to have different labels in the reflectance estimate. Notice that the weights always verify  $\alpha \leq \beta \leq \gamma$  and  $\alpha \ll \gamma$ .

with the intensity of its neighbouring regions, causing unwanted intensity edges in the resulting shading image. We address this global coherence problem by modifying the intensity of the the RGB descriptors of each uniform reflectance area according to the intensity of the original image. To solve this inference problem we use a belief propagation algorithm.

Let  $I = \bigcup_{i \in \mathcal{U}} U_i$  be the partition of an image into its areas of uniform reflectance provided by the MRF. Let  $L_i$  and  $L_i^{orig}$  be the intensities of the RGB triplet of the area  $U_i$  and of the same area in the original image, respectively. Ideally, to reflect the real shading, the ratio of intensities should verify, for each pair of areas in contact  $U_i$  and  $U_j$ ,

$$\frac{L_i}{L_j} = \frac{L_i^{orig}}{L_j^{orig}}. \quad (3.8)$$

However, the connectivity between uniform reflectance areas is complex and usually there is no transformation able to make all the intensity ratios similar to those of the original image in general. In practice, we minimize the differences using the mean squared error (MSE). In this minimization problem, we want regions sharing a long boundary to have a higher weight. The length of the boundary between two regions (denoted  $l_{ij}$  for regions  $i$  and  $j$ ) is defined to be the amount of pixels in both regions which have a neighboring pixel (assuming 4-neighborhood) belonging to the other region.

Thus, our purpose is to find a set of scalars  $\{\lambda_i | i \in \mathcal{U}\}$  which modify our estimated reflectance intensities in order to enhance the global coherence of our recovered shading scene (explicitly,  $L_i$  is substituted by  $\lambda_i L_i$ ). Mathematically, we can define a function  $W$  depending on such scalars as

$$W(\{\lambda_i\}_{i \in \mathcal{U}}) = \sum_{(i,j) \in \mathcal{U}^2, i < j} l_{i,j} \|\lambda_i L_i L_j^{orig} - \lambda_j L_j L_i^{orig}\|_2 \quad (3.9)$$

and find the set of values that minimize it:

$$\{\lambda_i\}^* = \arg \min_{\{\lambda_i\}} W(\{\lambda_i\}). \quad (3.10)$$

This can be done by applying the MSE and imposing a lower bound to the solution (otherwise we could obtain the trivial solution  $\lambda_i = 0, \forall i \in \mathcal{U}$ ). Figure 3.7 shows an example of how global coherence strongly improves the accuracy of the method, resulting in better intrinsic estimates.

### 3.5 Experiments

In this section we evaluate the performance of our approach. First, we recall the error metrics for intrinsic image evaluation that have been proposed in previous works. Afterwards, we test our method on the MIT dataset [71], which has become the standard set to test intrinsic image algorithms. We quantitatively and qualitatively compare our results to the ones obtained by several previous approaches.

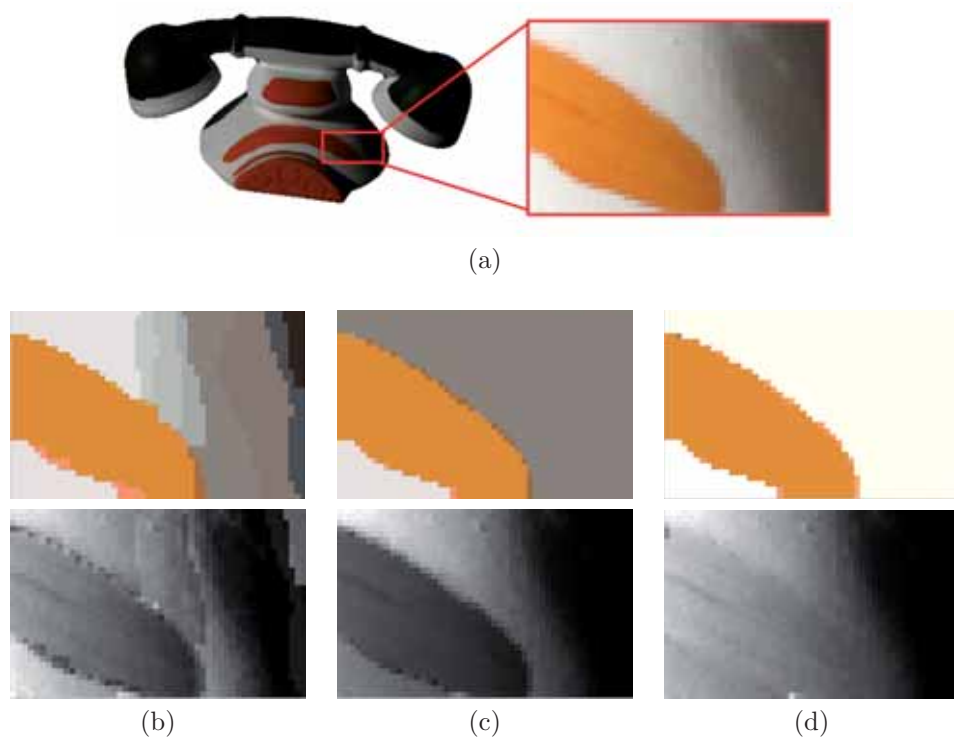


Figure 3.7: (a) Original Image. Different reflectance and shading estimates are shown: (b) Only local information of the color-name descriptor is used. (c) Output of the MRF where semi-local information of the ridge observations has been combined with local information of the color-name descriptor. (d) Adding global scene coherence to the output of the MRF.

As we have seen in Chapter 2, several error metrics have been proposed in previous works to evaluate intrinsic image algorithms. One of the most used metrics in previous works is the local MSE (LMSE), which was proposed by Grosse *et al.* [71] as an appropriate measure for edge-based methods. The authors claimed that, for such methods, the MSE is too restrictive because images with just a small misclassified edge can have a large MSE. However, Jiang *et al.* [90] argued that the LMSE sometimes has a low value in images that are not qualitatively good. To overcome this problem they defined a new metric based on the LMSE, the absolute LMSE (aLMSE), and also proposed to evaluate intrinsic images using the correlation measure, which computes the similarity, *i.e.* statistical dependency, between two images independently of their mean values.

Whereas correlation and the MSE are global error measures, the others are variations of the global measures and are computed as an average of local error on small image windows. In this work we have observed that the LMSE is biased towards edge-based methods. Hence, in the next section, although we evaluate our method

with four error metrics, we will focus our analysis on the results for global measures, such as the MSE and correlation.

We have estimated the intrinsic images of the MIT-20 dataset, composed of 20 images whose reflectance and shading ground truth are available. For each image, the error has been obtained by averaging the results on the reflectance and shading estimates. We compared our results to state-of-the-art approaches<sup>†</sup>. Previous methods have been evaluated either on the original MIT-16 dataset, composed of 16 scenes, or on the extended MIT-20 dataset, which includes 4 extra scenes. In each case, we compare our results with those from the methods whose results, or the code to generate them, are available.

The results on the MIT-16 dataset have been compared to grey and color Retinex algorithms (obtained from [71]), the methods by Tappen *et al.* (Tap-05 [146] and Tap-06 [145]), the methods by Shen and Yao [138] (Shen-SR and Shen-SRC), and Weiss' algorithm [157]. For the comparison on the MIT-20 dataset we test grey and color Retinex algorithms, the methods by Jiang *et al.* [90] (Jiang-A, Jiang-H and Jiang-HA), and Weiss' algorithm. The error metrics used to evaluate the results are the global measures MSE and correlation, and the local measures LMSE and aLMSE.

In our method, we set  $\mu = 1/3$  to weigh the two components of the energy function of our MRF and the dependence relations among parameters  $\alpha$ ,  $\beta$ , and  $\gamma$  defined in Section 3.3.2 as follows:  $\gamma/\alpha = 100$  and  $\gamma/\beta = 2$ . To initialize the network, we apply a logarithm to the input image before obtaining its color-shade descriptor.

Tables 3.1 and 3.2 show the results obtained by the evaluated methods on the MIT-16 and MIT-20 datasets, respectively. As can be seen in the tables, our method obtains the best results on the global measures (*i.e.* correlation and the MSE) on both the MIT-16 and the MIT-20 datasets when compared to single-image methods. In both cases, we overcame the results of the other methods, obtaining even better results than Weiss' algorithm, which uses image sequences and, therefore, has more information than single-image based methods. Notice that for the three error metrics, the lower is the better, while for the correlation the opposite holds. Observe that grayscale reflectances (computed as the mean of the RGB channels of the reflectance images) have been used in our evaluation in order to provide results which can be compared with the other published results.

As expected, local measures (the LMSE and the aLMSE) penalize our results and the performance of our methods considerably decreases when evaluated with these measures. However, as stated above, we consider that the evaluation of intrinsic image algorithms in terms of correlation or the MSE is more accurate since these measures are more meaningful in terms of similarity to the ground truth.

In Table 3.3, we present some qualitative results of our method on three objects of the MIT dataset. For comparative purposes, the objects shown are the ones used in [71]. These objects belong to each of the three subgroups of objects that the dataset contains, namely painted objects, printed papers, and animals. We compare our results to the ones from Color Retinex and Weiss algorithm, which are the best methods in the evaluation done in [71], and the results of the SRC method of Shen and Yao [138], which was the state-of-the-art<sup>‡</sup>.

---

<sup>†</sup>at the time of publication of [133] (2012)

<sup>‡</sup>at the time of publication of [133] (2012)

Method	Global measures		Local measures	
	Corr.	MSE	LMSE	aLMSE
Grey Retinex [71]	0.6494	0.1205	0.0329	0.3373
Tappen05 [146]	—	—	0.0570	—
Tappen06 [145]	—	—	0.0390	—
Col. Retinex [71]	0.7146	0.1108	0.0286	0.2541
Shen-SR [138]	0.7259	0.1223	0.0242	0.2454
Shen-SRC [138]	0.7733	0.0906	<b>0.0149</b>	<b>0.2147</b>
Ours [133]	<b>0.7862</b>	<b>0.0834</b>	0.0340	0.2958
Weiss [157]	0.7709	0.0900	0.0210	<b>0.1953</b>

Table 3.1: Results on the MIT-16 dataset (16 objects) with different error metrics. Shen-SRC results are computed on a subset of 13 objects ('deer', 'squirrel' and 'dinosaur' results were not available).

Method	Global measures		Local measures	
	Corr.	MSE	LMSE	aLMSE
Grey Retinex [71]	0.6292	0.1169	0.0296	0.3789
Col. Retinex [71]	0.7171	0.1072	<b>0.0257</b>	<b>0.2895</b>
Jiang-A [90]	0.6262	—	0.0388	0.4036
Jiang-H [90]	0.6179	—	0.0409	0.3655
Jiang-HA [90]	0.6631	—	0.0460	0.3655
Ours [133]	<b>0.7556</b>	<b>0.0836</b>	0.0305	0.3457
Weiss [157]	<b>0.7619</b>	0.0890	<b>0.0191</b>	<b>0.2230</b>

Table 3.2: Results on the MIT-20 dataset (20 objects) with different error metrics.

As can be seen in the table, our method is the only one that completely avoids the cast shadow on the reflectance image of the raccoon. Although the final colors of surfaces in the reflectance are not well recovered in the turtle's reflectance image, our method forces a single reflectance value within the areas where the material color is uniform and all the shading effects due to the textured surface of the shell are correctly included in the shading image. Such a result is a consequence of the color-shade descriptor, which enhances the stability of color names through illuminations variations. Finally, on the tea bag most of the errors are found on the shading estimate, which includes some reflectance information. However, the reflectance image is quite well recovered.

For the sake of completion we include our results for all the images in the MIT-16 and MIT-20 datasets. Quantitative results for different metrics can be seen in Tables 3.4 and 3.5, while qualitative results are presented in Tables 3.6, 3.7, 3.8 and 3.9.

We also provide the individual error values for each image on the MIT dataset with the four measures we have introduced in this section. For each image, we provide its error on the shading and the reflectance images, as well as the global averaged error. Boldface values in the tables correspond to the mean error values presented in Tables 3.1 and 3.2.

IMAGE			
GT			
COLOR RETINEX	 0.7533	 0.7873	 0.3325
SHEN SRC	 0.8625	 0.8454	 0.4558
OURS	 0.9687	 0.8835	 0.7020
WEISS	 0.5701	 0.9697	 0.7453

Table 3.3: Shading and reflectance images recovered by previous algorithms and our approach from images of the MIT database. Values below each decomposition are the corresponding correlation measures.



Image	MSE			Correlation		
	Shading	Reflectance	Global	Shading	Reflectance	Global
box	0,1749	0,2534	0,2142	0,8188	0,5567	0,6878
cup1	0,0044	0,0321	0,0183	0,9934	0,8675	0,9305
cup2	0,0683	0,1001	0,0842	0,8644	0,6934	0,7789
deer	0,0983	0,1099	0,1041	0,7313	0,8271	0,7792
dinosaur	0,0823	0,0411	0,0617	0,8145	0,9512	0,8829
frog1	0,1781	0,2051	0,1916	0,7094	0,7005	0,7050
frog2	0,1326	0,1480	0,1403	0,7217	0,2263	0,4740
panther	0,0055	0,0130	0,0092	0,9860	0,8267	0,9064
paper1	0,0028	0,0034	0,0031	0,9922	0,8868	0,9395
paper2	0,0064	0,1122	0,0593	0,9848	0,2639	0,6243
raccoon	0,0047	0,0043	0,0045	0,9931	0,9443	0,9687
squirrel	0,1674	0,1876	0,1775	0,6346	0,8061	0,7204
sun	0,0250	0,0148	0,0199	0,9153	0,9594	0,9374
teabag1	0,0864	0,0745	0,0804	0,4646	0,8574	0,6610
teabag2	0,0440	0,0491	0,0465	0,8222	0,9448	0,8835
turtle	0,1966	0,0412	0,1189	0,7238	0,6802	0,7020
apple	0,1587	0,1600	0,1593	0,8469	0,6785	0,7627
pear	0,0865	0,1158	0,1011	0,8213	0,0315	0,4264
phone	0,0575	0,0195	0,0385	0,8566	0,9742	0,9154
potato	0,0409	0,0379	0,0394	0,8781	-0,0204	0,4288
Mean 16 obj.	0,0799	0,0869	<b>0,0834</b>	0,8231	0,7495	<b>0,7863</b>
Mean 20 obj.	0,0811	0,0862	<b>0,0836</b>	0,8287	0,6828	<b>0,7557</b>

Table 3.4: MSE and correlation results on the MIT-16 and MIT-20 datasets.

Finally, Figure 3.1 shows how our method works with a natural image which has been previously used by other authors [32, 138].

### 3.6 Discussion and conclusion

In this chapter we have described a method for intrinsic image estimation based on two color cues. The first cue is based on the semantic description that human beings use to describe color values. A MRF has been used to combine this sparse description of color with a color-shade attribute based on an analysis of the color distribution of the image in the histogram space. This attribute enhances the stability of color names against strong changes in the illumination due to shadows and highlights. We have shown that a color-name descriptor based on psychophysical data provides a good basis for describing object reflectance.

Our results show how using color information in the problem of intrinsic image decomposition results in better estimates of the intrinsic reflectances of images. In agreement with other methods, we have assumed that the scenes are illuminated with a single “white light”, and the effects of the camera sensors used to acquire the input images have been ignored. In Chapter 4 we will discuss the influence of these effects

Image	LMSE			aLMSE		
	Shading	Reflectance	Global	Shading	Reflectance	Global
box	0,0314	0,0714	0,0514	0,3764	0,1693	0,2728
cup1	0,0019	0,0159	0,0089	0,1179	0,4285	0,2732
cup2	0,0492	0,0326	0,0409	0,4420	0,5026	0,4723
deer	0,0559	0,0467	0,0513	0,5306	0,3658	0,4482
dinosaur	0,0402	0,0201	0,0302	0,3055	0,1165	0,2110
frog1	0,0439	0,0815	0,0627	0,2854	0,5330	0,4092
frog2	0,0479	0,0390	0,0434	0,2364	0,5910	0,4137
panther	0,0029	0,0035	0,0032	0,0509	0,0829	0,0669
paper1	0,0021	0,0022	0,0021	0,0721	0,1238	0,0980
paper2	0,0038	0,0203	0,0121	0,1584	0,1609	0,1597
raccoon	0,0035	0,0035	0,0035	0,0383	0,1170	0,0776
squirrel	0,0804	0,0848	0,0826	0,6089	0,4593	0,5341
sun	0,0088	0,0035	0,0062	0,1539	0,0575	0,1057
teabag1	0,0650	0,0523	0,0587	0,6490	0,2138	0,4314
teabag2	0,0329	0,0355	0,0342	0,3013	0,1772	0,2393
turtle	0,0773	0,0278	0,0525	0,2084	0,8307	0,5196
apple	0,0202	0,0199	0,0200	0,3582	0,9178	0,6380
pear	0,0144	0,0171	0,0157	0,5782	0,7957	0,6869
phone	0,0096	0,0080	0,0088	0,1560	0,0714	0,1137
potato	0,0239	0,0202	0,0221	0,4882	0,9979	0,7430
Mean 16 obj.	0,0342	0,0338	<b>0,0340</b>	0,2835	0,3081	<b>0,2958</b>
Mean 20 obj.	0,0308	0,0303	<b>0,0305</b>	0,3058	0,3856	<b>0,3457</b>

Table 3.5: LMSE and aLMSE results on the MIT-16 and MIT-20 datasets.

in the problem of intrinsic image decomposition, how they affect existing methods and how they can be included in the formulation of the problem.


























Original Object	Estimated Shading	Estimated Reflectance	GT Shading	GT Reflectance
				
				
				
				
				

Table 3.6: Qualitative results for the “animal” images in the MIT dataset.










Original Object	Estimated Shading	Estimated Reflectance	GT Shading	GT Reflectance
				
				

Table 3.7: Qualitative results for the “printed paper” images in the MIT dataset.














































Original Object	Estimated Shading	Estimated Reflectance	GT Shading	GT Reflectance
				
				
				
				
				
				
				
				
				

Table 3.8: Qualitative results for the “painted object” images in the MIT dataset.





















Original Object	Estimated Shading	Estimated Reflectance	GT Shading	GT Reflectance
				
				
				
				

Table 3.9: Qualitative results for the 4 extra images in the MIT-20 dataset.



# Chapter 4

## A General Framework Based on the Photometry of Intrinsic Images

In this chapter we examine the inaccuracy of existing intrinsic image formulations to properly account for the effects of illuminant color and sensor characteristics in the estimation of intrinsic images and present a generic framework which incorporates insights from color constancy research to the intrinsic image decomposition problem. The proposed mathematical formulation includes information about the color of the illuminant and the effects of the camera sensors, both of which contribute to the observed reflectance of the objects in a scene during the acquisition process. By modeling these effects, we get a “truly intrinsic” reflectance image (we call it absolute reflectance), which is invariant to changes of illuminant or camera sensors. This framework allows us to represent a wide range of intrinsic image decompositions depending on the specific assumptions on the geometric properties of the scene configuration and the spectral properties of the light source and the acquisition system, thus unifying previous formulations in a single general framework.

### 4.1 Motivation

Intrinsic image decomposition methods initially focused on providing reflectance and shading image estimates of a scene [27, 157, 57]. Moreover, other subfields in computer vision estimate different intrinsic characteristics. For example, shape from shading methods [45] estimate the shape (*i.e.* orientation, depth...) of the objects given a shading image, color constancy methods [79, 63] estimate the illuminant of the scene, and highlight removal techniques [13] estimate image specularities.

As we have mentioned in Chapter 1, intrinsic characterization of scenes is fundamental in multiple computer vision applications. The complex ways in which light interacts with shapes and materials continue to confound solutions in areas that range from 3D shape reconstruction to object recognition and to material identification. In this chapter we argue that in order to increase accuracy in such applications multiple intrinsic properties have to be studied in conjunction so that the appropriate intrinsic image for each property can be estimated.

We have already seen in chapter 2 that early methods of intrinsic image estimation either worked with grayscale images [104, 27, 157] or assumed scenes with Lambertian surfaces (*i.e.* surfaces reflecting the same amount of light in all directions) and a single white light source [146, 57], thus simplifying the decomposition into intrinsic components to the product of its intrinsic images of shading and reflectance. This simplified formulation has been commonly used ever since (*e.g.* [90, 161, 136, 138, 60, 133, 37]). In [25], Beigpour and van de Weijer relaxed the Lambertian assumption and added a specular term to the formulation. Specularity detection is a hard problem in itself and several specular removal techniques can be found in the literature (see [13] for a survey). Lately, in [19, 22], Barron and Malik relaxed the white light assumption while keeping the Lambertian surfaces and single light source constraints. In their work, the shading image was modeled as a function of the shape of the objects in the scene, which can be described in many ways (depth maps, normal maps, etc.), and the color and geometry of the light source of the scene.

All these formulations are consistent with the physically based dichromatic reflection model [135], which describes how light is reflected in a scene under some simplifying assumptions. However, the color value at each image pixel cannot be determined using only the reflection model. During the image acquisition process, three factors influence the color value that we finally measure at each pixel: the reflectance of the objects in a scene, the illuminant of the scene, and the spectral response of the camera sensors. While reflectance is an intrinsic property of objects, the illuminant and the camera sensors modify the way we see this reflectance in the images. That is why we must study and describe separately these three factors by isolating the effects of both illuminant and camera sensors. So far, however, existing methods on intrinsic image decomposition provide reflectance intrinsic images which contain mixed information about the color of the illuminant (which is assumed white in many approaches), object reflectance values and camera sensor effects. To the best of our knowledge camera sensors have not been studied before in the area of intrinsic image decomposition.

We draw insights from the rich literature in color constancy (hitherto somewhat disconnected from intrinsic image research), which also aims to a stable representation of object color across different images. From the start, the aim of color constancy [53] has been to estimate the color of a scene under a canonical light source, which is similar to the problem of reflectance estimation. Multiple works [103, 51, 52, 149, 79, 63] have provided different methods to estimate the scene illuminant from a single image, which in turn allows to remove its effects and obtain a canonical color image. Research in color constancy has also devoted some attention to the effects of the imaging sensors for the final recovery of the canonical illuminants [152]. The narrow band property of the spectral sensitivity of the sensors simplifies the characterization of changes in the illuminant [48]. Therefore, the subsequent estimation of the color of the illuminant is also simplified. In this work we build upon the connection between the fields of color constancy and intrinsic image decomposition by extending the intrinsic model to include the effects of the color of the light source and the biases introduced by the sensors. Both factors have direct effects on the computation of the reflectance.

Moreover, camera sensors also affect the image values in the acquisition process. Usually cameras have three sensors (*i.e.* RGB sensors) which filter incoming light



at different wavelengths and transform it into electrical signals by means of a set of predefined sensitivity functions. Each camera model has different sensors with different sensitivity functions. Therefore, the pixel values we observe in the images are also dependent on the camera sensors. Usually, the sensitivity functions are not specified. However, some camera calibration models estimate the sensitivity functions of the camera [16, 124, 153, 93, 41], allowing the removal of sensor effects from images.

In the formulation we propose, it is possible to study the effects of the camera sensors and the color of the illuminant and isolate them from the reflectance image. As a result, we obtain a new reflectance image which is invariant to changes of the camera sensors and the color of the illuminant. This reflectance image, which we call absolute reflectance, is a truly intrinsic image. The invariance of our absolute reflectance image provides many practical advantages in different scenarios. Our method is practically applicable because such sensor information is publicly available for many sensors and the color of the light can be estimated with any state of the art color constancy method. As we demonstrate in Section 4.4.1, even partial knowledge of sensor characteristics leads to significant improvement in the estimated intrinsic images.

## 4.2 Reflectance and color fundamentals

In this section we introduce some theoretical aspects about the image acquisition process. We observe the role of the color of the illuminant and the camera sensors in the acquisition process and also analyze how these effects have been modeled in the field of computer vision.

### 4.2.1 Color image formation: physics

When the light source can be assimilated to a point source, the spectral radiance outgoing from an infinitesimal patch at location  $x$  on an object  $obj$  along the direction  $v_r$  can be physically expressed as

$$L_{obj}(x, v_r, \lambda) = \int_{\Omega_x} f_r(x, v_r, v_i, \lambda) L_i(\lambda) \cos(\theta_i) d\omega_i. \quad (4.1)$$

In this expression [86],  $L_i(\lambda)$  is the incident radiance (*i.e.* irradiance),  $\cos(\theta_i)$  captures the geometry of the scene by expressing the reduction in the amount of light impinging the surface at  $x$  due to the angle  $\theta_i$  between the surface normal (*i.e.* normal to the patch) and the incident direction  $v_i$ , and  $f_r(x, v_r, v_i, \lambda)$  is the bidirectional reflectance distribution function (brdf) of the object that specifies how much of the incident light coming from direction  $v_i$  is reflected into the viewing direction  $v_r$  per unit wavelength  $\lambda$  at  $x$ . Finally,  $\Omega_x$  denotes the hemisphere given by the normal to the patch. Figure 4.1 illustrates the variables used in Equation 4.1. The brdf captures how light and material interact to shape the appearance of an object. In particular, the spectral power of the light outgoing from an object depends on both the spectrum of the light source and the reflectance of the object.

Several reflection models have been proposed in the field of computer vision and computer graphics [121] which assume different simplifications of the physics of light.

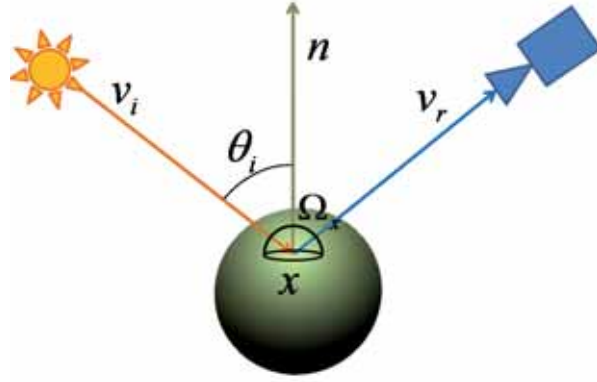


Figure 4.1: Illustration of the multiple variables that model the spectral radiance.

Here we use Shafer's Dichromatic Reflection Model (DRM) [135] because it offers a good trade-off between realism and practicability.

In Shafer's DRM, light reflection is modeled as two separate reflection processes, each having a characteristic spectral power distribution whose magnitude varies with the direction of illumination and viewing. In the DRM, the amount of light reflected per unit wavelength in direction  $v_r$  by a small surface patch at a point  $x$  of an object  $obj$  is defined as

$$L_{obj}(x, v_r, \lambda) = \int_{\Omega_x} (m_b(x, v_r, v_i)c_b(\lambda) + m_s(x, v_r, v_i)c_s(\lambda)) d\omega_i, \quad (4.2)$$

where  $c_b$  and  $c_s$  are functions describing the spectral distributions of the body reflectance and the surface reflectance of the object, respectively, and  $m_b$  and  $m_s$ , called diffuse and specular magnitudes, are geometric factors that weigh the amount of light coming from the body and the surface of the object, respectively, and only depend on the geometry of the scene. Equation 4.1 and the DRM in Equation 4.2 are equivalent when the brdf function,  $f_r$ , is defined in the following way:

$$f_r(x, v_r, v_i, \lambda) = h_b(x, v_r, v_i)g_b(x, \lambda) + h_s(x, v_r, v_i), \quad (4.3)$$

where the scalar  $h_b(x, v_r, v_i)$  describes the variations in body reflection due to the geometry of the scene,  $g_b(x, \lambda)$  describes the reflection of the body at  $x$ , which is supposed to be isotropic (*i.e.* independent of  $v_r$  and  $v_i$ ), and the scalar  $h_s(x, v_r, v_i)$  describes the variations in surface reflection depending on the geometry of the scene. When we substitute the right-hand side of Equation 4.3 in Equation 4.1, we obtain

$$L_{obj}(x, v_r, \lambda) = \int_{\Omega_x} (h_b(x, v_r, v_i)g_b(x, \lambda)L_i(\lambda) \cos(\theta_i) + h_s(x, v_r, v_i)L_i(\lambda) \cos(\theta_i))d\omega_i. \quad (4.4)$$

We now want to see that Equations 4.2 and 4.4 are equivalent. The diffuse and specular magnitudes are functions that only depend on the geometry of the scene.

Therefore, we can write

$$m_b(x, v_r, v_i) = h_b(x, v_r, v_i) \cos(\theta_i) \quad (4.5)$$

and

$$m_s(x, v_r, v_i) = h_s(x, v_r, v_i) \cos(\theta_i). \quad (4.6)$$

The spectral power distribution of the body of the object depends on that of the incident light,  $L_i(\lambda)$ , as well as the reflectance function of the body of the object,  $g_b(x, \lambda)$ , which is independent of  $x$  when uniformly colored objects are assumed as in the DRM, thus

$$c_b(\lambda) = g_b(\lambda)L_i(\lambda). \quad (4.7)$$

Finally, the spectral power distribution of the surface reflectance of the object is that of the incident light,

$$c_s(\lambda) = L_i(\lambda). \quad (4.8)$$

To sum up, at the physical level the spectral composition of the light reflected from objects in a scene depends on the intrinsic reflectance of the object and of the spectrum of the light. Moreover, this interdependency can be modeled mathematically, which will prove to be useful for the formulation of the intrinsic decomposition below.

### 4.2.2 Color image formation: sensors

As explained above, the light reflected from an object depends on both its material reflectance and the color of the light. Additionally, in the acquisition process, the resulting image values are also affected by a third factor, namely the camera sensors.

Cameras use a finite set of sensor responses to describe the continuous light spectrum. Camera sensors vary widely with the characteristics of the camera, and different cameras usually produce different measurements. The values measured by the sensors of a camera are obtained by spectral integration [97],

$$p_k = \int_{\Lambda} L_{obj}(x, v_r, \lambda) S_k(\lambda) d\lambda, \quad (4.9)$$

where  $S_k(\lambda)$ ,  $k = 1, 2, 3$ , are functions describing the absorption curves of the camera sensors and  $L_{obj}(x, v_r, \lambda)$  is the reflected light from object, *obj*, in a scene.  $\Lambda$  denotes the integration domain, which corresponds to the visible spectrum. The dependence of  $p_k$  on  $x$  is further omitted for the sake of simplicity.

For most computer vision problems, it is important to transform the measurements  $p_k$  made with a given camera to measurements made with standard sensors  $S_k^s(\lambda)^*$ ,

$$p_k^s = \int_{\Lambda} L_{obj}(x, v_r, \lambda) S_k^s(\lambda) d\lambda. \quad (4.10)$$

There exist several computational models of camera calibration that estimate the sensitivity functions of the camera [16, 124, 153, 93, 41]. Given an estimate of the sensitivity functions of a camera, and standard sensitivity functions, a transformation

---

\*  $s$  superscript stands for “standard”.

matrix  $\mathbf{S}_{Sen}$  converts theoretical responses for standard sensors into those of the actual camera. Sensor transformations are often described by 3-by-3 matrices [85]. If  $\mathbf{p}^s$  denotes the vector whose standard coordinates are  $p_k^s$ , and  $\mathbf{p}$  denotes the response of the camera whose coordinates are  $p_k$ , we have

$$\mathbf{p} = \mathbf{S}_{Sen} \cdot \mathbf{p}^s. \quad (4.11)$$

Notice, however, that the term  $\mathbf{S}_{Sen}$  is usually an approximation of the sensor transformation. In this work we adopt the widely used standard RGB (sRGB) sensitivity curves [7] as the standard sensor.

Moreover, as we will see in the next section, a valuable property for a set of camera sensors is to be narrow band. We say that a set of camera sensors is narrow band when the overlap among the spectral responses of these sensors is small, meaning that the response for each of the sensors scarcely influences the responses of the other sensors. Narrow band responses are uncorrelated, which proves to be critical for most computer vision applications, among which color constancy [50].

### 4.2.3 Joint illumination/sensor modeling for intrinsic images

As we have seen in the previous section, both the spectral distribution of the light source and the sensor properties affect camera measurements. Defining color constancy algorithms able to provide image representations invariant to these dependencies has been a long-standing goal in the computer vision community. Originally, the term color constancy refers to the human ability to maintain a stable color representation of a scene irrespective of its illuminant.

A classical approach in computational color constancy is to estimate the color of the illuminant of a scene using a single image and then subtract the illuminant to build a stable image under a canonical illuminant (see [79, 63] for a survey). The canonical representation of the color of an image is given by

$$p_k^c = \int_{\Lambda} L_{obj}^c(x, v_r, \lambda) S_k(\lambda) d\lambda, \quad (4.12)$$

where  $L_{obj}^c(x, v_r, \lambda)^\dagger$  is the light reflected by the object under the canonical illuminant. In this work we use the CIE standard illuminant D65 as the canonical illuminant [131].

Let us now describe illuminant and sensor transformations in a practical way. In the color constancy literature, 3-by-3 matrices have been commonly used to describe illuminant transformations [53]. We denote by  $\mathbf{L}_{CLig}$  the color conversion that transforms pixel values under the canonical illuminant to values under the actual illuminant. Therefore, we have

$$\mathbf{p} = \mathbf{L}_{CLig} \cdot \mathbf{p}^c. \quad (4.13)$$

Notice that in the last equation the term  $\mathbf{L}_{CLig}$  is usually an approximation of the illuminant transformation. Several methods have been proposed to find the appropriate transformation matrix  $\mathbf{L}_{CLig}$  [63] and most of them rely on the sharpness of the

---

<sup>†</sup><sub>c</sub> superscript stands for “canonical”.

sensors and approximate the illuminant using a diagonal model [158, 48]. Spectral sharpening methods [50] provide sensor transformations,  $\mathcal{T}$ , that convert a given set of sensor sensitivity functions into a new set of functions which are less overlapped,

$$\mathcal{T} \cdot \mathbf{p} \approx \mathbf{L}_{CLig} \cdot \mathcal{T} \cdot \mathbf{p}^c. \quad (4.14)$$

In general, spectral sharpening improves the performance of color constancy algorithms that are based on an independent adjustment of the sensor response channels [50, 152].

Our objective is to describe the intrinsic components by discarding both the effects of the color of the light and the dependence on a particular set of camera sensors. The values

$$p_k^{s,c} = \int_{\Lambda} L_{obj}^c(x, v_r, \lambda) S_k^s(\lambda) d\lambda \quad (4.15)$$

represent the appearance of the objects under a canonical illuminant and standard camera sensors. Hence, from previous statements, we derive that the relationship between  $\mathbf{p}^{s,c}$  and  $\mathbf{p}$  is

$$\mathbf{p} = \mathbf{L}_{CLig} \cdot \mathbf{S}_{Sen} \cdot \mathbf{p}^{s,c}. \quad (4.16)$$

In this equation,  $\mathbf{L}_{CLig}$  and  $\mathbf{S}_{Sen}$ , isolate lighting and sensor effects, respectively. As we have already mentioned before, these terms are usually just good approximations of the illuminant and sensor transformations respectively. However, the errors of such approximations are ignored in this work for the sake of simplicity.

### 4.3 A general model for intrinsic image estimation

Thus far, in the literature the photometric issues described above have been disregarded. Moreover, some assumptions have been consistently made to simplify the problem. Under the assumptions that surfaces are Lambertian and that there is a single white light in the scene, the intrinsic image decomposition problem is reduced to the estimation of shading and reflectance images [27, 157, 146, 57] according to

$$I(x, y) = I_{Shad}(x, y) \cdot I_{Refl}(x, y), \quad (4.17)$$

where  $I_{Shad}$  represents the amount of reflection arriving to the point  $(x, y)$  of the image from a specific point of the object surface considering the shape of the objects in the scene and the position of the light source, and  $I_{Refl}$  describes how the light is reflected by the corresponding point of the object considering the material reflectance properties. Since the model is assumed to be defined for any image point, for the sake of simplicity the notation  $(x, y)$  is further omitted.

The Lambertian assumption was relaxed in [25], where the authors added a specular term to the model

$$I = I_{Shad} \circ^{\ddagger} I_{Refl} + I_{Spec}, \quad (4.18)$$

where  $I_{Spec}$  denotes the specular reflection of the objects in the scene.

---

<sup>‡</sup>Here 'o' denotes the Hadamard product.

Barron and Malik [19] modeled the shading image as a function,  $M$ , of the shape of the scene,  $I_{Shape}$ , and the color and direction of the illuminant,  $\mathbf{L} = [\mathbf{L}_{CLig}, \mathbf{L}_{GLig}]$ , which led to the decomposition

$$I = I_{Shad} \circ I_{Refl} = M(I_{Shape}, \mathbf{L}) \circ I_{Refl}. \quad (4.19)$$

Although Barron’s attempt to model the illumination of the scene in their approach of intrinsic image estimation, the effects of camera sensors and illumination have not been jointly considered. As explained above, image values are affected by both factors and, therefore, the reflectances recovered by previous methods depend on the illuminant and the sensor used to acquire the image. We refer to such recovered reflectances as relative reflectances.

To overcome the dependence problems that relative reflectance images have, in this chapter we propose a general framework for intrinsic image decomposition which takes into account both illuminant and camera sensor effects, thus allowing us to recover the reflectance images as if they were acquired with standard sRGB sensors under the canonical illuminant. We define such images as absolute reflectance images, and denote them  $I_{Refl}^a$ .

As introduced in Equation 4.16 the physical properties of the scene at a pixel (reflection model, geometry, etc.) encoded in  $\mathbf{p}^{s,c}$  can be isolated from the effects of the camera sensors and the illuminant of the scene, described respectively in  $\mathbf{S}_{Sen}$  and  $\mathbf{L}_{CLig}$ . According to the DRM (Equations 4.2 and 4.9), we can decompose these values,  $\mathbf{p}^{s,c}$ , into their diffuse and specular components as

$$\mathbf{p}^{s,c} = \mathbf{p}_{Shad}^{s,c} \circ \mathbf{p}_{Refl}^{s,c} + \mathbf{p}_{Spec}^{s,c}, \quad (4.20)$$

where  $\mathbf{p}_{Shad}^{s,c}$  and  $\mathbf{p}_{Refl}^{s,c}$  describe the magnitude and composition of the body (*i.e.* diffuse) reflection, and  $\mathbf{p}_{Spec}^{s,c}$  represents the magnitude of the surface (*i.e.* specular) reflection. Hence, from Equations 4.16 and 4.20 the general formulation at a pixel,  $\mathbf{p}$ , is

$$\mathbf{p} = \mathbf{L}_{CLig} \cdot \mathbf{S}_{Sen} \cdot (\mathbf{p}_{Shad}^{s,c} \circ \mathbf{p}_{Refl}^{s,c} + \mathbf{p}_{Spec}^{s,c}). \quad (4.21)$$

Equation 4.21 can be extended from a pixel level to a whole image level, leading to our proposal for a general framework for intrinsic image estimation which deals with absolute reflectances:

$$I = \mathbf{L}_{CLig} \cdot \mathbf{S}_{Sen} \cdot (I_{Shad} \circ I_{Refl}^a + I_{Spec}). \quad (4.22)$$

Our proposed framework, illustrated in Figure 4.2, models the effects of the light source and the camera sensors and includes a novel reflectance term which is invariant to these effects.

### 4.3.1 Model particularities and relation to previous models

Depending on the knowledge we have about the scene and the acquisition conditions, or the assumptions we make on them, our model leads to different simplifications of the general formula proposed in Equation 4.22. Furthermore, existing formulations (Equations 4.17-4.19) in the intrinsic image decomposition field can be derived from

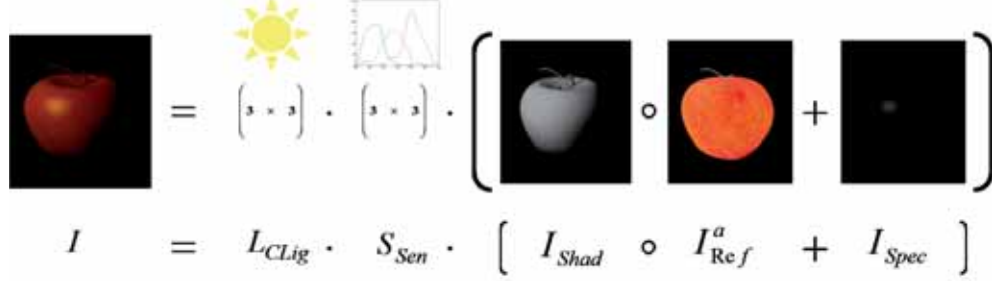


Figure 4.2: Overview of the proposed model. Modeling the effects of the illumination and the sensors in the acquired images allows intrinsic image methods to obtain absolute reflectance images  $I_{Ref}^a$ .

our model as specific cases. Let us describe some of these assumptions and discuss how they influence the general formulation of our model. In our formulation of the problem the effects of the color of the light source and the response of the camera sensors are modeled using 3-by-3 matrices. The resulting reflectance image, or absolute reflectance, is invariant to changes in the illuminant of the scene or the camera sensors.

When surfaces in a scene are Lambertian, specularities can be discarded from the general formulation (i.e.  $I_{Spec} = 0$ ) and Equation 4.22 becomes:

$$I = \mathbf{L}_{CLig} \cdot \mathbf{S}_{Sen} \cdot (I_{Shad} \circ I_{Ref}^a). \quad (4.23)$$

When there is a single canonical light source, the illuminant transformation happens to be a diagonal matrix with a single scalar value  $\kappa$  for the three channels. This transformation modifies the general intensity of the image. Therefore,  $\mathbf{L}_{CLig}$  is substituted by the scalar value  $\kappa$ :

$$I = \kappa \cdot \mathbf{S}_{Sen} (I_{Shad} \circ I_{Ref}^a + I_{Spec}). \quad (4.24)$$

When camera sensors are narrow band,  $\mathbf{L}_{CLig}$  can be described by a diagonal matrix. This illuminant transformation matrix can be represented as an image,  $\mathbf{L}_{CLig} = I_{CLig}$ , where each channel in  $I_{CLig}$  contains one value of the diagonal matrix  $\mathbf{L}_{CLig}$ . Under this notation, the Hadamard product can be used in order to multiply this term with the shading and the absolute reflectance components in the intrinsic image decomposition:

$$I = \mathbf{S}_{Sen} \cdot (I_{CLig} \circ I_{Shad} \circ I_{Ref}^a + I_{CLig} \circ I_{Spec}). \quad (4.25)$$

As explained in Section 4.2.3, when the sensitivity curves of the camera sensors are overlapped, spectral sharpening methods can be applied to make them more narrow band. This allows us to write  $\mathbf{L}_{CLig} = I_{CLig}$  as in the previous case. Including the transform  $\mathcal{T}$  defined in Equation 4.14, the formulation above stands as:

$$I = \mathcal{T}^{-1} \cdot [I_{CLig} \circ I_{Shad} \circ (\mathcal{T} \cdot \mathbf{S}_{Sen} \cdot I_{Ref}^a) + I_{CLig} \circ (\mathcal{T} \cdot \mathbf{S}_{Sen} \cdot I_{Spec})]. \quad (4.26)$$

When the camera sensors are the canonical sRGB sensors,  $\mathbf{S}_{Sen}$  is the identity matrix, and Equation 4.22 results in

$$I = \mathbf{L}_{CLig} \cdot (I_{Shad} \circ I_{RefI}^a + I_{Spec}). \quad (4.27)$$

Notice that even when we only have partial information about light and sensors, we are still able to estimate the illuminant transformation matrix (by applying color constancy methods) and the camera sensor behaviour (sometimes available on public websites), which makes our method widely applicable.

However, when the conditions described above are not fulfilled, as it is usually the case with natural images, we cannot accurately estimate the absolute intrinsic reflectance  $I_{RefI}^a$ . In such case, the absolute reflectance can only be approximated by a relative reflectance image,  $I_{RefI}^r$ , which includes part of the illumination and sensor effects. For instance, when we have no knowledge about the camera sensors and canonical sRGB sensors are assumed (for example, when estimating  $\mathbf{S}_{Sen}$  is not possible), the estimated reflectance is  $I_{RefI}^r = \mathbf{S}_{Sen} \cdot I_{RefI}^a$ .

Existing models for intrinsic image decomposition can also be derived from this generalized model by considering their specific assumptions. Several approaches [146, 145, 133] assume Lambertian objects ( $I_{Spec} = 0$ ), white light ( $\mathbf{L}_{CLig} = \kappa$ ) and unknown camera sensors ( $\mathbf{S}_{Sen} = Id$ ), which yields

$$I = \kappa \cdot (I_{Shad} \circ I_{RefI}^r), \quad (4.28)$$

where  $I_{RefI}^r = \mathbf{L}_{CLig} \cdot \mathbf{S}_{Sen} \cdot (I_{RefI}^a + I_{Spec})$ . We assume that the methods that do not consider the specular component will include highlights in the reflectance image. This last equation is equivalent to Equation 4.17 up to a scale factor. Other approaches [25] deal with the specular image term while keeping the white light assumption ( $\mathbf{L}_{CLig} = \kappa$ ) and having no knowledge of the camera sensors ( $\mathbf{S}_{Sen} = Id$ ). This leads to

$$I = \kappa \cdot (I_{Shad} \circ I_{RefI}^r + I_{Spec}), \quad (4.29)$$

where  $I_{RefI}^r = \mathbf{L}_{CLig} \cdot \mathbf{S}_{Sen} \cdot I_{RefI}^a$ . This, in fact, is Equation 4.18 up to a scale factor. Finally, Barron et al.[19] (Equation 4.19) relaxed the white light assumption. If narrow band sensors are assumed ( $\mathbf{L}_{CLig} = I_{CLig}$ ) our model becomes:

$$I = I_{Shad} \circ I_{CLig} \circ I_{RefI}^r, \quad (4.30)$$

where  $I_{RefI}^r = \tilde{\mathbf{L}}_{CLig} \cdot \mathbf{S}_{Sen} \cdot (I_{RefI}^a + I_{Spec})$ , and  $\tilde{\mathbf{L}}_{CLig}$  represents the information of the light source not modeled by  $I_{CLig}$  when the sensors are not narrow-band.

In Table 4.1, we show how the proposed model can be simplified by usual assumptions such as images acquired under the canonical light, scenes with only Lambertian surfaces, or cameras with narrow-band sensors. Afterwards, in Table 4.2, we show how the formulations of the decomposition in intrinsic components used in previous methods are particular cases of our extended formulation. These particular formulation depend on the specific assumptions of each method.



Formulation	Assumptions
$I = \mathbf{L}_{CLig} \cdot \mathbf{S}_{Sen} \cdot (I_{Shad} \circ I_{Ref}^a + I_{Spec})$	None
$I = \mathbf{L}_{CLig} \cdot \mathbf{S}_{Sen} \cdot (I_{Shad} \circ I_{Ref}^a)$	Lambertian surfaces $I_{Spec} = 0$
$I = \kappa \cdot \mathbf{S}_{Sen} \cdot (I_{Shad} \circ I_{Ref}^a + I_{Spec})$	Canonical/White light $\mathbf{L}_{CLig} = Id \cdot \kappa$
$I = \mathbf{S}_{Sen} \cdot (I_{Shad} \circ I_{CLig} \circ I_{Ref}^a + I_{CLig}^{(*)} \circ I_{Spec})$	Narrow-band sensors diagonal $\mathbf{L}_{CLig}$
$I = \mathcal{T}^{-1} \cdot [I_{Shad} \circ I_{CLig} \circ (\mathcal{T} \cdot \mathbf{S}_{Sen} \cdot I_{Ref}^a) + I_{CLig} \circ (\mathcal{T} \cdot \mathbf{S}_{Sen} \cdot I_{Spec})]$	Sharpened sensors diagonal $\mathbf{L}_{CLig}$
$I = \mathbf{L}_{CLig} \cdot (I_{Shad} \circ I_{Ref}^a + I_{Spec})$	Standard sensors (sRGB) $\mathbf{S}_{Sen} = Id$

(\*) When  $\mathbf{L}_{CLig}$  is a diagonal matrix, it may be replaced by an image  $I_{CLig}$ . This allows us to turn the matrix product into an image product. In such case, this image  $I_{CLig}$  directly represents the color of the light source.

Table 4.1: Specific formulations of the model resulting from common assumptions.

Previous works	Assumptions	Specific formulation	Estimated reflectance
Tappen <i>et al.</i> [145] Shen <i>et al.</i> [137] Jiang <i>et al.</i> [90] Shen <i>et al.</i> [136] Shen-Yeo [138] Gehler <i>et al.</i> [60] Serra <i>et al.</i> [133]	Lambertian Canonical light Standard sensors	$I = \kappa \cdot (I_{Shad} \circ I_{Ref}^r)$	$I_{Ref}^r = \mathbf{L}_{CLig} \cdot \mathbf{S}_{Sen} \cdot (I_{Ref}^a + I_{Spec}^{(*)})$
Beigpour <i>et al.</i> [25]	Canonical light Standard sensors	$I = \kappa \cdot (I_{Shad} \circ I_{Ref}^r + I_{Spec})$	$I_{Ref}^r = \mathbf{L}_{CLig} \cdot \mathbf{S}_{Sen} \cdot I_{Ref}^a$
Barron-Malik [19]	Lambertian Narrow-band sensors	$I = I_{Shad} \circ I_{CLig} \circ I_{Ref}^r$	$I_{Ref}^r = \tilde{\mathbf{L}}_{CLig}^{(**)} \cdot \mathbf{S}_{Sen} \cdot (I_{Ref}^a + I_{Spec})$

(\*) We assume that the methods that do not consider the specular component will include highlights in the reflectance image.

(\*\*) The matrix  $\tilde{\mathbf{L}}_{CLig}$  represents the information of the light source which is not modeled by  $I_{CLig}$  when the sensors are not narrow-band.

Table 4.2: Summary of how previous methods for intrinsic image decomposition are related to our general model.

## 4.4 Validation Experiments

In the previous sections we have proposed a new model for intrinsic image decomposition that revisits the simplistic image product formulation. Our approach includes constraints derived from the consideration of the color of the light source and the bias introduced by the acquisition process. In this section we present a set of experiments to validate the proposed framework. Our experiments simulate multiple scenarios which differ by the degree of knowledge of the scene. These scenarios involve different assumptions on the sensors, the light source, or the illuminant model:

- The camera sensors may be either known (in which case we use the corresponding sensor responsivity functions), or unknown (*i.e.* standard sRGB sensors are assumed).
- The illuminant of the scene may be either known (in which case we use the corresponding spectral power distribution), or unknown (*i.e.* a canonical D65 illuminant is assumed).
- The model of illumination can either be diagonal (*i.e.* sharp sensors are assumed) or a full 3-by-3 model.

Taking these considerations into account, we define four scenarios (summarized in Table 4.3) which describe the most common situations found in real applications:

*Scenario 1 (SC1):* It corresponds to the ideal case. Both camera sensors and scene light source are known, and  $\mathbf{L}_{CLig}$  is approximated by a full illuminant model.

*Scenario 2 (SC2):* It is identical to SC1 except that a diagonal illumination model is assumed, just as most color constancy methods do.

*Scenario 3 (SC3):* Here the camera sensors are unknown (*i.e.* standard sensors are assumed) but the scene illuminant is known or can be estimated with a diagonal model.

*Scenario 4 (SC4):* Here no knowledge of either the camera sensors (*i.e.* standard sensors are assumed) or of the lighting conditions (*i.e.* a canonical illuminant is assumed) is available.

We use the angular error to calculate the differences between the estimated reflectance values in each scenario and the ground truth reflectance values. The angular error between two RGB values,  $\mathbf{p}_i$  and  $\mathbf{p}_j$ , is defined as the angle between the vectors determined by these values,

$$A_e(\mathbf{p}_i, \mathbf{p}_j) = \cos^{-1} \left( \frac{\mathbf{p}_i \cdot \mathbf{p}_j}{\|\mathbf{p}_i\| \cdot \|\mathbf{p}_j\|} \right). \quad (4.31)$$

This error metric has been previously used in the literature to test the accuracy of color constancy methods [79, 63].

Scenario	Camera Sensors	Scene Illuminant	Illuminant Model
SC1	Known	Known	Full
SC2	Known	Known	Diagonal
SC3	Unknown	Known	Diagonal
SC4	Unknown	Unknown	-

Table 4.3: Description of the scenarios used in the experiments.

#### 4.4.1 Experiment 1: Synthetic data

For this experiment we use a dataset of 1995 reflectances compiled from several sources in Barnard *et al.* dataset [17]. These reflectances include the 24 Macbeth color-checker patches, 1269 Munsell chips, 120 Dupont paint chips, 170 natural objects, the 350 surfaces in Krinov dataset, and 57 additional reflectances. These data allow us to synthesize the scenes for the experiment. To simulate the light source of the scenes, we have used the spectral power distributions of 10 illuminants: 5 Planckian and 5 non-Planckian (see Appendix B). These illuminants have been chosen from the same dataset (Barnard *et al.* [17]). To simulate the imaging process we have used 7 sets of commercial cameras sensors (Canon EOS1D, Sigma Foveon D10, Kodak DCS420, Leica M8, Nikon D70, Olympus E400, Sony DXC930, and TVI MSC-1024RGB12).

We have performed the experiments on 10 scenes, each one synthesized with 140 different reflectances randomly chosen from Barnard’s dataset. Each scene has been integrated with each of the 10 illuminants and subsequently with each of the 7 sets of camera sensors, providing the pixel values  $\mathbf{p}$  for the 700 resulting scenes. Our ground-truth (*i.e.* absolute reflectances) are the same 10 scenes integrated with the canonical D65 illuminant spectrum and standard sRGB sensitivity curves. The ground-truth values are denoted by  $\mathbf{p}^{s,c}$ .

In order to estimate the transformation matrices  $\mathbf{S}_{Sen}$  and  $\mathbf{L}_{CLig}$ , we also synthesize relative reflectance values by integrating each reflectance with the spectrum of the canonical illuminant D65 and each of the 7 sets of camera sensors, resulting in a set of  $\mathbf{p}^c$  values. We approximate the camera sensor transformation matrix,  $\mathbf{S}_{Sen}$ , from  $\mathbf{p}^{s,c}$  and  $\mathbf{p}^c$  using a 3-by-3 matrix estimated by least squares minimization. The illuminant transformation matrix  $\mathbf{L}_{CLig}$  can be estimated from  $\mathbf{p}$  and  $\mathbf{p}^c$  values following the same procedure.

Figure 4.3 illustrates the results of this experiment. We observe that when some knowledge about the camera sensors and the illuminant of the scene is available (SC1), the error is considerably reduced with respect to the worst case scenario, where we do not have any further information of the scene and a D65 illuminant and sRGB sensors are assumed (SC4). On average, the error in scenario 4 is reduced by 93% when knowledge about the illuminant and the camera sensors is available (SC1). For the most extreme case, the Nikon D70 camera, the error is reduced by 97.5%. Most computer vision techniques have so far ignored the effects of the illuminant of the scene and the camera sensors, including these errors into their results.

When we approximate the effect of the illuminant using a diagonal model (SC2),

the errors increase by 48% on average with respect to scenario 1. Nonetheless, the errors in scenario 2 are always smaller than 5 degrees. These results are coherent with previous results showing that diagonal models are sufficient to correctly describe illuminant transformations [48].

When we only have partial information about the scene and we estimate the illuminant of the scene but assume sRGB sensors (SC3), for most cameras the errors also decrease notably with respect to the worst case scenario (SC4). On average, the errors in scenario 3 are reduced by 42% with respect to the errors in scenario 4. However, there are three cameras (Sigma Foveon D10, Kodak DCS420 and Nikon D70) for which the errors decrease by less than 10%. These error differences within cameras for scenario 3 are probably due to the response of the camera sensors. Cameras with sensor sensitivity functions similar to sRGB curves get much larger error reductions than other cameras, since sRGB sensors are assumed in scenario 3.

The numerical results of this experiment can be seen in Table 4.4. The results are presented separately for Planckian and non-Planckian illuminants in order to study possible differences in the behavior of our framework which result from the type of illumination in the scene. We observe that, in general, errors are slightly larger when non-Planckian illuminants are used. However, the behavior of our framework with Planckian and non-Planckian illuminants is exactly the same.

The conclusion we draw from this experiment is that any knowledge about the illuminant of the scene and the camera sensors, even if it is only partial knowledge, is critical for many computer vision tasks involving color images. Specifically, in intrinsic image decomposition it means that the more information we have about the camera sensors and the scene illuminant, the better our reflectance estimate will be.

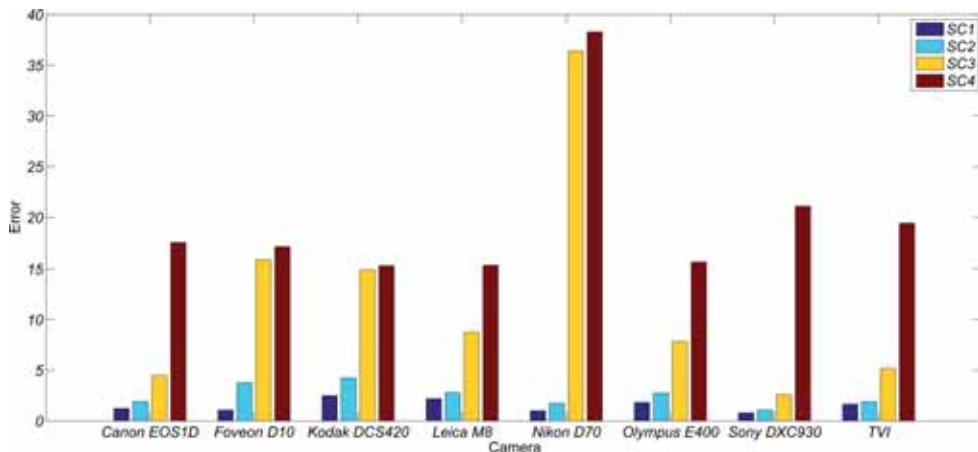


Figure 4.3: Mean angular error for different commercial cameras in multiple scenarios. The most information we have about the scene, the best our reflectance estimate will be. Scenario 4 always assumes canonical illumination and standard sensors, like most existing intrinsic image models. Scenarios 1 and 2 represent our model when there is some knowledge about scene illumination and camera sensors.

Camera	Planckian				Non-Planckian			
	SC1	SC2	SC3	SC4	SC1	SC2	SC3	SC4
Canon EOS1D	1.19	1.55	4.47	12.93	1.26	2.27	4.51	22.27
Sigma Foveon D10	1.02	2.74	15.45	16.41	1.10	4.78	16.34	17.96
Kodak DCS420	2.45	3.68	14.35	14.99	2.53	4.80	15.43	15.61
Leica M8	2.14	2.48	8.55	12.50	2.23	3.11	8.97	18.19
Nikon D70	0.95	1.36	36.28	37.67	1.05	2.12	36.50	38.89
Olympus E400	1.69	2.29	7.59	12.47	1.93	3.21	8.04	18.86
Sony DXC930	0.79	1.04	2.67	15.33	0.80	1.09	2.48	26.84
TVI	1.76	1.93	5.45	16.16	1.53	1.88	4.87	22.67

Table 4.4: Numerical results (angular error in degrees) of the experiments showed in Figure 4.3.

#### 4.4.2 Experiment 2: Natural images

In our second experiment, we use a set of images from Flickr [3]. We choose images from six well-known scenes (Figure 4.4), taken with four popular DSLR camera models, namely the Canon EOS 5D Mark II, the Nikon D50, the Nikon D7000 and the Sony NEX-5.



Figure 4.4: Image categories: From left to right, the Golden Gate bridge (San Francisco), the Statue of Liberty (New York City), the Kinkakuji temple (Kyoto), the church of Sagrada Família (Barcelona), the Uluru Rock (Australian desert) and St. Basil's cathedral (Moscow).

The images selected for this experiment fulfill the basic assumptions of our model: scenes with a single light source and known camera sensors. The amount of pictures that we choose for each camera model and image category depends on the number of available pictures which satisfy these requirements, but on average we have 4 pictures per camera model and scene. The total number of images used in our experiment is 107. We first remove the gamma correction, which we assume to be 2.2, from all the pictures. Then, for each picture, we select two or three regions which describe a single material under similar illumination conditions (*i.e.* overshadowed and saturated areas of the picture are avoided). The mean value of these regions is used as a color descriptor for this picture.

For each of the pictures we apply two color constancy methods (Shades of Grey [52] and Grey-Edge [149]) and remove the effects of the sensors for each of them (*i.e.* express the image values under sRGB sensors) by using the transformation matrices obtained from DxOMark website [2].

Figure 4.5 shows some qualitative results where the improvement after applying a color constancy algorithm and removing the effect of the camera sensors is clear. In Figure 4.6 we present quantitative results using a color constancy algorithm to estimate the illuminant of the scenes. The color descriptors for each image are compared with the color descriptors of the other images representing the same scene category using the chromatic angular error. We also average the obtained error between pair of pictures and observe that this error decreases accordingly to the available information about the illuminant and the camera sensors. When the effects of both the illuminant and the camera sensors are removed, there is an average decrease in the error of 22.59%. These results are coherent with those we obtain in our previous experiment with synthetic data. Although the error reduction is more significant when we use synthetic data, in both experiments the results show that the error decreases accordingly to the amount of available information about the illuminant of the scene and the camera sensors. The only exception is found in the results for the Nikon D7000 camera, where the error in scenario 3 (illuminant known, sensors unknown) is bigger than the error in scenario 4 (illuminant and sensors unknown). However, when both the illuminant and camera sensors are known (SC2) the error for this camera also decreases. Furthermore, all the images on the second experiment have been acquired under different daylight illumination conditions (*i.e.* Planckian illuminants). On the other hand, the diversity of the illuminants used in the first experiment is much bigger, since they are defined in a bigger space, including both Planckian and non-Planckian illuminants. As a result, the decrease of the errors in the experiment with synthetic data is more pronounced than it is in the experiment with natural images.

We have seen how the illuminant of the scene and the camera sensors affect the resulting image values, but our aim is to show how isolating these effects can benefit the performance of many intrinsic image decomposition methods. Figure 4.7 shows how our model can be applied in practice to the problem of intrinsic image estimation. Given pairs of images describing the same scene under different illumination conditions and taken with different camera models, the reflectance estimates for Serra *et al.* method [133] are much closer when the effects of the illuminant and the camera sensors are removed. For the first example (Sagrada Família), the angular error of the estimated reflectance decreases by 63.4%, while the error for the second example (Uluru Rock) decreases by 82.3%.

### 4.4.3 Experiment 3: Laboratory-acquired images

In this subsection we include experiments on synthetically rendered and real lab-acquired images that complement the results exposed in the previous experiments.

In Figures 4.8 and 4.9, we show the effects of considering the photometric properties through the proposed model prior to applying usual methods for estimating reflectance intrinsic images.

In Figure 4.8 we show results on a synthetic specular object which has been rendered simulating different illuminants and different camera sensors (column (a)). The scenes in Figure 4.9 (original images in column (a) in both figures) have been acquired in a laboratory with a common sensor and under different illuminants (placing

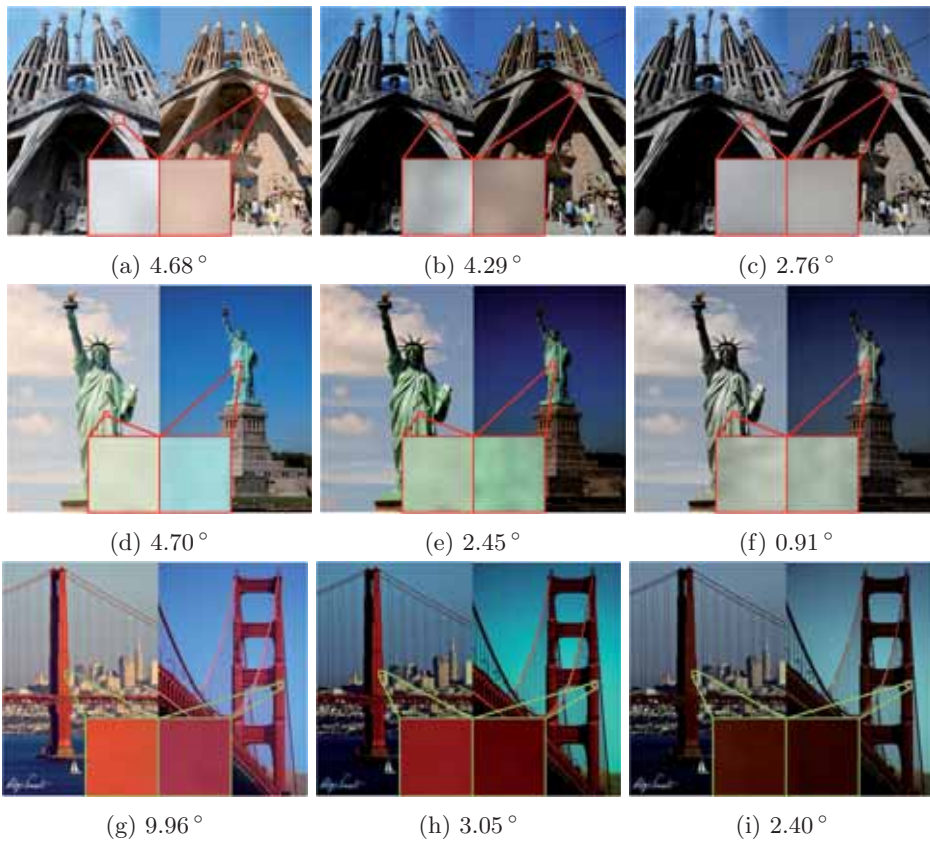


Figure 4.5: Results on 3 pairs of natural images. At each row, in the first column we see a pair of original images. In the second column the Grey-Edge algorithm [149] has been applied. In the last column, we see the images of the second column after their camera sensor effects have been removed. The chromatic angular errors are expressed in degrees.



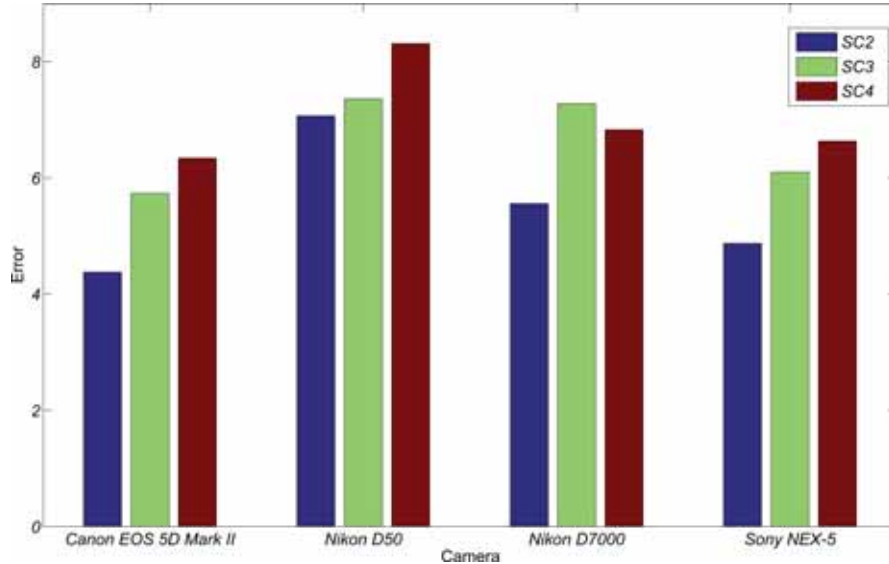


Figure 4.6: Mean angular error among pairs of images from our set of natural images corresponding to the scenarios 2, 3 and 4 previously defined. The Shades of Grey algorithm [52] has been used to estimate the illuminant, but similar results are obtained when using other color constancy methods, such as Grey-Edge [149].

different color filters in front of our light source).

For each image acquired in the laboratory we captured an extra picture containing a Macbeth color-checker. This enabled us to approximate the transformation matrix for the illuminant,  $L_{CLig}$ , and the transformation matrix for the camera sensors,  $S_{Sen}$ , by means of a least squares algorithm. The matrices  $S_{Sen}$  and  $L_{CLig}$  were used to transform input images into more stable images, which were closer to an image of the scene acquired under the canonical illuminant and with standard sensors.

For all three objects we show the reflectance images estimated by Serra *et al.* [133] and Barron *et al.* [20] in columns (b) and (c), respectively. Subsequently, in columns (d) and (e) we show the estimated reflectances computed by the same methods after the effects of both the color of the illuminant and the camera sensors have been removed from input images.

In all scenes we calculate the mean angular error, averaged over all the pixels of the object, between the different pairs of estimated reflectance images. Notice that the error decreases for both methods when we remove the effects of the color of the illuminant and the camera sensors from input images, showing that the use of photometric information increases the stability of the recovered reflectance.

Finally, in Figure 4.10 we show a Lambertian object with multiple color values on its surface. We have acquired images of this object under different illumination conditions (column (a)). In order to measure the convenience of estimating an absolute reflectance image instead of a relative reflectance image, we first project on the chromatic plane the relative and absolute images that we get for each illuminant. Then,

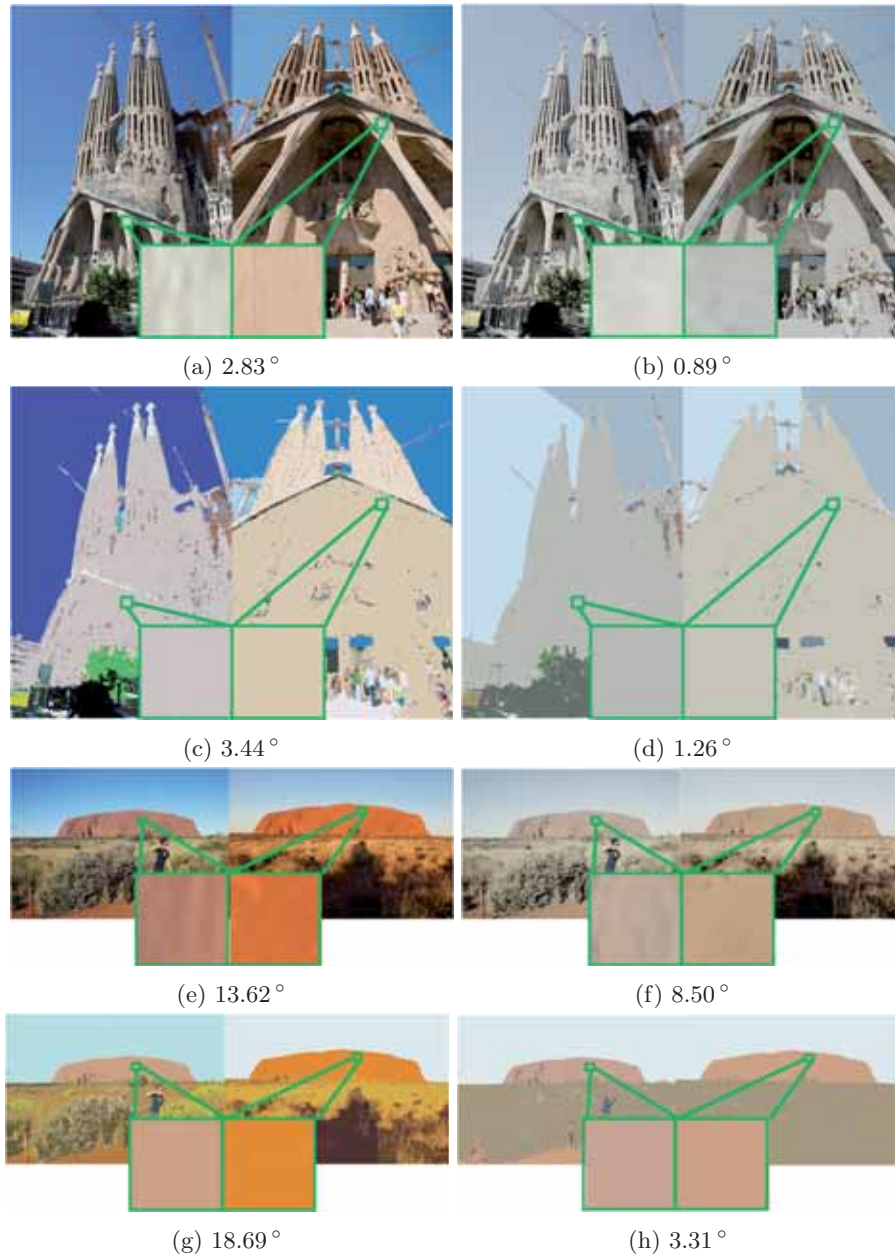


Figure 4.7: Example of the influence of the color of the illuminant and the camera sensors on intrinsic images. (a) and (e) are images of a given landmark, taken with different cameras under different illumination conditions. (b) and (f) are the images in (a) and (e) respectively, after removing the effects of the illuminant and the camera sensors. (c),(d),(g) and (h) are the estimated intrinsic images of (a),(b),(e) and (f), respectively. Chromatic angular error values are given in degrees.

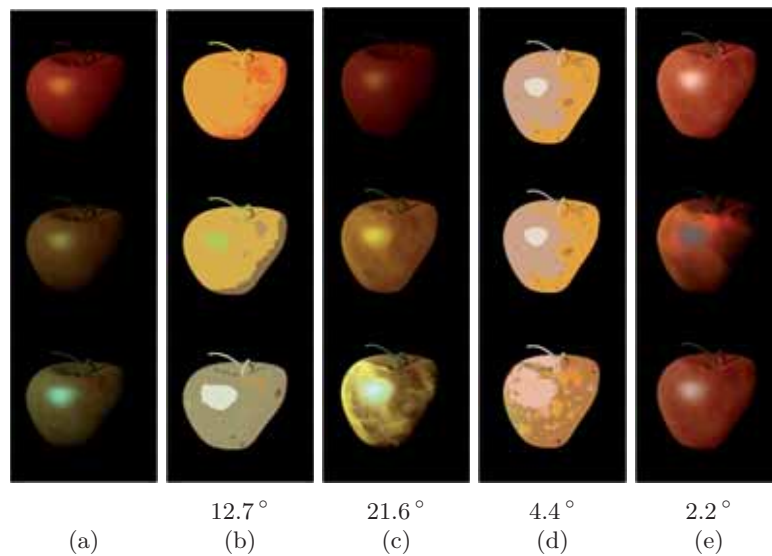


Figure 4.8: Results showing the average chromatic difference for a synthetic object rendered under different lighting and sensor conditions. The first and second rows share the same camera sensors but not lighting conditions. The second and third rows share the scene illuminant but not the camera sensors. Ideally all 3 images should have the same intrinsic reflectance image. Numbers under each column are the average difference between images as measured by angular differences in sRGB space. (a) Original Image. (b) Reflectance estimates of the method of Serra *et al.* [133]. (c) Reflectance estimates of the method of Barron *et al.* [19]. (d) and (e) Removing the effects of the color of the illuminant and the camera sensors from input images minimizes the differences between the three reflectance estimates for [133] and [19], respectively.

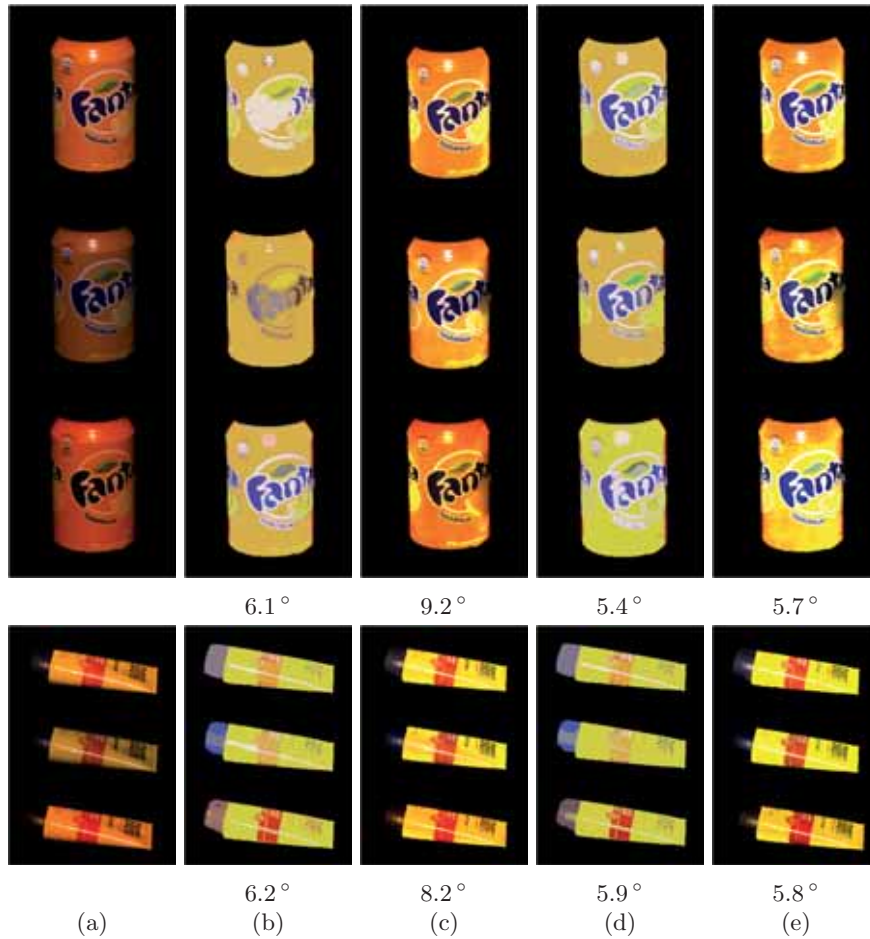


Figure 4.9: Results showing the average chromatic difference for two objects acquired in the laboratory under different lighting conditions. Ideally, for each object, all 3 images should have the same intrinsic reflectance image. Numbers under each column are the average chromatic difference between images as measured by angular differences in sRGB space. (a) Original Image. (b) Reflectance estimates of the method of Serra *et al.* [133]. (c) Reflectance estimates of the method of Barron *et al.* [19]. (d) and (e) Reflectance estimates for [133] and [19], respectively, after removing the effects of the color of the illuminant and the camera sensors.

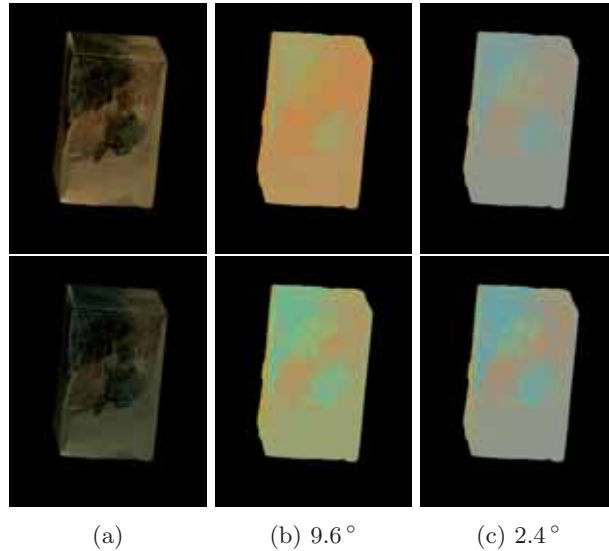


Figure 4.10: Variability of relative and absolute reflectance of an object under two different illuminants. (a) Original image. (b) Theoretical relative reflectance. (c) Theoretical absolute reflectance.

we measure the difference between the two relative reflectances and the difference between the two absolute reflectances using the mean angular error. In column (b) we show the relative reflectances of the object under the two illuminants projected on the chromatic plane, while the corresponding absolute reflectances, projected on the same plane, are shown in column (c). Since the angular difference between the ground truth reflectance components of both images decreases considerably when the effects of the color of the illuminant and the camera sensors are removed from input images, we conclude that the absolute reflectance (column (c)) is far more an “intrinsic” characteristic than the relative reflectance (column (b)) that has been commonly used so far in the literature.

## 4.5 Discussion and conclusion

The main contribution of this chapter is a new theoretical framework for intrinsic image decomposition which includes the photometric effects of the illuminant of the scene and the camera sensors. This model generalizes the previous formulations of the problem. Another sound contribution of this work is the definition of absolute reflectances, which are invariant to the color of the light source and the response of the camera sensors.

As a direct consequence of our framework we can model camera effects. When the camera sensors are sufficiently narrow band, the illuminant transformation model is described by means of a diagonal matrix. When sensor responses are not narrow band but known, spectral sharpening [50] can be applied. Otherwise, a full (*i.e.* non-

diagonal) linear model is used. Furthermore, we can model scene illumination. This is an important step towards the normalization of color through different acquisitions with the same device. We believe that for widely used sensors such as the Kinect, such corrections could even be incorporated in the standard libraries for a wide use by developers in any application that relies on accurate matching of surface appearances across different images.

Absolute reflectances could prove very useful to many applications. Imagine a set of cameras placed outdoors with known camera sensors (*e.g.* transit cameras placed along a highway). Illumination effects will be very diverse depending on the location of each camera and the time of the day and the year when the images were taken. However, if we estimate the color of the illuminant using color constancy, then we could infer an absolute reflectance image which would be very useful for further computer vision tasks. Imagine we have multiple camera devices in a lab. We do not know anything about their sensors, but we have control over the scene illumination and know its spectral distribution. If we place an object with some known reflectance values in the scene, we can estimate a sensor transformation for each camera. This way, we can estimate absolute reflectance images for any camera. Finally, imagine we have a camera with unknown sensors and we have no knowledge about the illuminant either. However, we know or can approximate some reflectance values which appear in the scene. This could lead to data-driven approaches for the inference of illuminant and sensor type. Furthermore, one could envision absolute reflectance descriptors, which would provide a characterization of object materials invariant to specific device characteristics.

# Chapter 5

## New Datasets for Intrinsic Image Evaluation

As we have previously seen in Chapter 2, building a dataset for intrinsic image estimation is challenging. The MIT dataset [71], for instance, 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 spraying the object with white paint to obtain the shading ground truth; polarizing filters are used to separate specular from Lambertian reflectance; and interreflections need to be avoided because they would lead to false ground truth images [71]. As a result, the MIT dataset presents some weaknesses: only single object scenes are present, all of them are captured under a white illuminant, more complex and realistic lighting conditions (*i.e.* multiple illuminants) are not considered, and interreflections are absent.

In [19], 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. Therefore, evaluation of intrinsic image methods needs new and more general datasets which also provide accurate ground truth information for multiple intrinsic characteristics of the scene also estimated by current methods (*e.g.* direction and color of the illuminant, shape, etc.).

In this chapter we present two new datasets for intrinsic image evaluation which aim to overcome some of the drawbacks of existing datasets. The first dataset includes ground truth information about the illuminant of the scene and the camera sensors. We will use this dataset to validate the general formulation for intrinsic image decomposition presented in the previous chapter. We will also show the influence of both the illuminant of the scene and the camera sensors in the problem of intrinsic image decomposition, remarking the need to model these factors in future methods for intrinsic image decomposition. The second dataset has been built using computer graphics. We will show how computer graphics can help us to easily build large ground truth collections which include realistic scenes with complex illumination effects.

## 5.1 A Calibrated Dataset for Intrinsic Image Estimation

In Chapter 4, we showed how the illuminant of a scene and the camera sensors affect the final color values in an image of the scene. We also proposed a novel framework for intrinsic image decomposition which models both factors and generalizes previous formulations.

In this section we present a calibrated dataset for intrinsic image evaluation which includes ground truth information about the illuminant of the scene and the sensors of the camera used in the acquisition process. Moreover, this dataset has been designed to be general enough to serve as an evaluation tool for most of the existing methods, as well as for future methods which fit in our general framework.

Our dataset contains 20 objects with diverse shapes, materials and textures (see Figure 5.1). These objects have been selected to provide a high variability of reflectances and thus encode different levels of complexity for decomposition methods. This dataset is available online\* to further motivate authors to model the effects of the color of the illuminant and the camera sensors in their approaches.



Figure 5.1: Objects in our Dataset.

### 5.1.1 Methods and Materials

We used the procedure adopted to build the MIT dataset [71] to acquire the images and recover accurate ground truth data. The lab in which we acquired the images of the dataset consists of a room of about 2m height by 3m width by 3m depth. The walls and the ceiling of the lab were covered with black felt to minimize interreflections, and the table in the center of the room was sprayed with black matte paint. Moreover, to avoid displacements of the objects during the ground truth acquisition process, we fixed the objects to a black platform located in the center of the table. We used the sphere fixation method described in [71], which allowed us to remove and replace the platform (and the object) in the exact same position.

\*[http://www.cic.uab.cat/Datasets/photometric\\_intrinsic\\_image\\_dataset](http://www.cic.uab.cat/Datasets/photometric_intrinsic_image_dataset)



In order to provide different illumination conditions in the dataset, three 2800K (*i.e.* orangish) bulb lights were placed at different positions in the lab. The different directions of the bulb lights already provide three illumination conditions. Two additional lighting conditions were set for each object by using two color filters (one blue, another yellow) which were placed in front of one of the bulbs. In total, we acquired each object under 5 illumination conditions consisting of different illuminant directions and colors. The lamps were fixed and remained untouched during the whole process and the color filters could be easily placed and removed using an independent system of hooks hanging from the ceiling. The colors of the light sources have been recovered using a ColorChecker Digital SG of 140 patches from X-Rite Photo [6], where the exact theoretical spectrum for each of its color patches is known. Moreover, the 3-by-3 illuminant transformation matrices, have been recovered using a least squares algorithm on images of the ColorChecker taken under the different illumination settings.

Finally, two cameras, namely a Nikon D5200 and a Sigma Foveon D10, were used. The sensor responses of the Sigma camera have been obtained from [124], while the ground truth information for the sensor effects of the Nikon camera have been obtained from the DxOMark photography website [2]. Both cameras present significant differences in their sensor responses. In addition, when acquiring images under the dim illumination of the lab, the Sigma camera provides challenging noisy images. Both cameras were fixed during the whole acquisition process and the positions of the objects, light sources and cameras were completely determined from the beginning. Figure 5.2 illustrates the distribution of the cameras and the light sources in our laboratory. The color filters that we used to modify the color of the illuminant are also represented in this figure as colored regions placed next to the light bulbs.

### 5.1.2 Acquisition Process

The procedure to acquire ground truth data for a given object was as follows: we first fixed the object to the platform and took three images with each camera (two cameras) for each illumination setting (five different settings). The images taken with the same camera under the same illumination conditions were used to recover a median image (where the value for each pixel  $(x, y)$  in the median image results from the median value of the pixels  $(x, y)$  in the three images). This median image decreases the amount of noise in our data, specially in the images captured with the Sigma camera. We then applied the illuminant and sensor transformation matrices to these images in order to remove the effects of the illuminant and the camera sensors. The transformed median images were denoted  $I^i$ , where  $i$  stands for the  $i^{th}$  lighting condition.

Then, we removed the platform with the object, painted the object white and replaced the platform using the fixation method mentioned above to avoid displacements. We captured the object again with both cameras for each lighting condition. We then applied the illuminant and sensor transformation matrices to these images. The transformed images were denoted  $I_{shad}^i$ , and their intensity images were used as shading ground truth.

Finally, in order to recover accurate absolute reflectance values,  $I_{refl}^a$ , and mini-

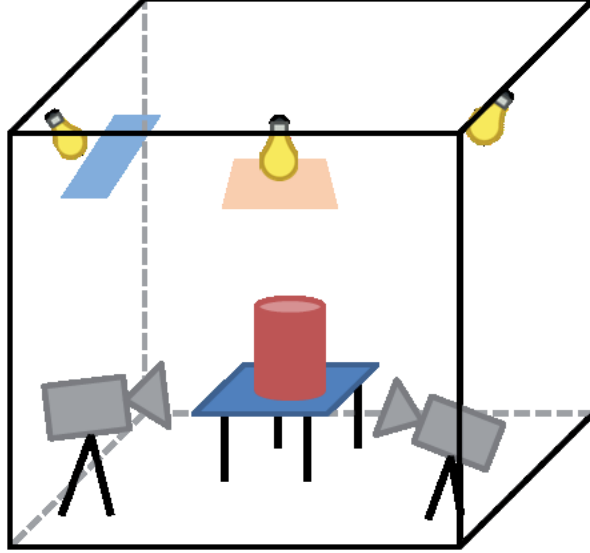


Figure 5.2: Schema of our laboratory. We illustrate the position of the two cameras, the three light bulbs and the object (in the center of the room). The color filters that we used to change the color of the illuminant are also represented, as colored regions, close to the light bulbs.

minimize the errors produced when dividing low values found in highly shadowed regions of the images, we used the equation

$$I_{refl}^a = \frac{\sum_{i \in Ldir} I^i}{\sum_{i \in Ldir} I_{shad}^i}, \quad (5.1)$$

where  $Ldir$  denotes the set of illumination conditions. In practice, we only used the first three illumination conditions (*i.e.* different light source directions and no color filter). Observe that we used multiple images of the same object taken under different light directions, as well as their white-painted counterparts. A similar process was also used in the MIT dataset [71] to recover the intrinsic reflectance images. However, in the MIT dataset white light was assumed and the effects of the camera sensors were ignored. As a consequence, the values of their reflectance images are influenced by both factors.

The ground truth data provided for a given object can be seen in Figure 5.3. The first row presents the absolute reflectance of the object and its binary mask. The next five rows represent different illumination conditions. In each row the original image is presented in the first column, while its respective shading component can be seen in the second column, and the third column provides a spherical representation of the illuminant of the scene, showing both its direction and color.

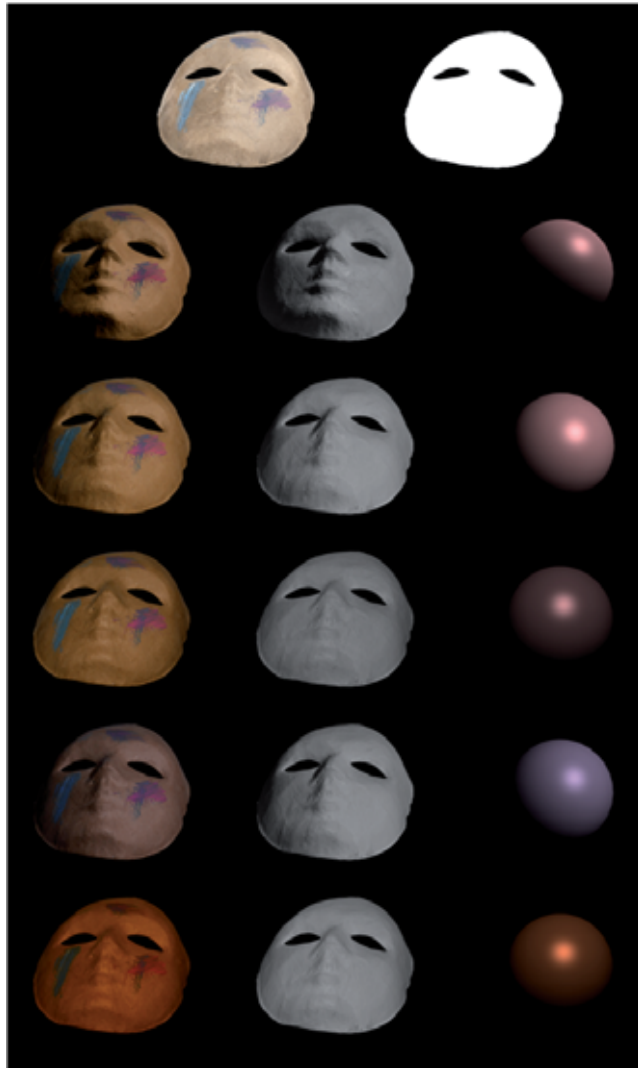


Figure 5.3: Example of the ground truth data we can provide for a single object and camera.

### 5.1.3 Experiments

In Chapter 4, we presented different experiments which demonstrated the influence that the color of the illuminant and the camera sensors have in the image values, and indicated that both factors affect the problem of intrinsic image decomposition. In this section we aim to provide quantitative results about the influence of the color of the illuminant and the camera sensors in the problem of intrinsic image estimation. To validate the proposed model in Chapter 4 we use different methods for intrinsic image decomposition, namely Jiang *et al.* [90], Gehler *et al.* [60], Serra *et al.* [133] and Barron and Malik [19]. We test these methods using the images in our dataset in three different scenarios, corresponding to some of the scenarios previously defined in Chapter 4:

*Scenario 1 (SC1)*: It corresponds to the ideal case. We know both camera sensors and scene light source. A full illumination model is assumed.

*Scenario 3b (SC3b)*: Here the camera sensors are unknown (*i.e.* standard sensors are assumed), but the scene illuminant is known. Observe that this scenario differs from scenario 3 because here the illuminant is estimated with a full model instead of a diagonal model.

*Scenario 4 (SC4)*: Here no knowledge of either the camera sensors (standard sensors are assumed) or the lighting conditions (a canonical illuminant is assumed) is available.

The local mean square error (LMSE) [71], is used to quantitatively evaluate the reflectance estimates for the different methods with respect to the ground truth absolute reflectance image. The LMSE has been questioned by some authors [90, 133]. However, it is still the most commonly used metric for intrinsic image evaluation.

The quantitative results of the evaluated methods, expressed in LMSE scores considering the three RGB channels of the color image, are presented in Table 5.1. These results show that the quality of the reflectance estimates is closely related to the amount of information available about the scene and the acquisition process. When information about the illuminant of the scene or the camera sensors is available, the performance of the methods improves (*i.e.* the errors are reduced) in all cases.

When we compare the performance of the methods when no information about the illuminant or the camera sensors is available (SC4) to the results they achieve when the effects of both factors are known or can be estimated (SC1), we see that, on average, the mean LMSE scores of the methods decrease by 56.38% and 40.12% for the cameras Nikon D5200 and Sigma Foveon D10, respectively. In particular, the mean LMSE error for the method of Gehler *et al.* decreases by 70.66% when tested with the Nikon D5200 images. Even for the method of Barron and Malik [19], which uses priors on scene illumination and jointly estimates the direction and color of the illuminant of the scene, we observe an impressive reduction of the LMSE when the effects of both the illuminant and the camera sensors are removed (67.48% and 50.59% for images acquired with the Nikon D5200 camera and the Sigma Foveon D10 camera, respectively). These results show the importance of modeling the effects of both the illuminant of the scene and the camera sensors in methods for intrinsic image

estimation. From these results, we can derive that existing methods which assume white light and sRGB sensors carry big errors in their intrinsic reflectance estimates.

However, some of the methods are more influenced by the effect of the color of the illuminant and the camera sensors than others. Observe, for example, that our method [134] clearly outperforms all the other methods when there is no available information about the illuminant of the scene or the camera sensors (SC4). Moreover, other methods achieve better results than our method when both factors are known (SC1).

Furthermore, notice that the mean LMSE errors for all methods are slightly higher for the images acquired with the Sigma Foveon D10 camera. This may be a result of the remaining noise in the images of the dataset. Although we reduced the noise of the images for both cameras, as we have explained in the previous section, some noise still remains, specially in the images acquired with the Sigma Foveon D10 camera. However, this noise does not affect dramatically the global performance of the different methods for intrinsic image decomposition. This is probably due to the fact that most methods for intrinsic image decomposition enhance the smoothness of the shading images and the sparsity of reflectance values.

	Nikon D5200			Sigma Foveon D10		
	SC1	SC3b	SC4	SC1	SC3b	SC4
Jiang <i>et al.</i>	0.059	0.077	0.131	0.086	0.095	0.154
Serra <i>et al.</i>	0.047	0.054	<b>0.070</b>	0.066	<b>0.065</b>	<b>0.077</b>
Gehler <i>et al.</i>	<b>0.033</b>	0.050	0.112	<b>0.057</b>	0.065	0.115
Barron & Malik	0.043	<b>0.047</b>	0.133	0.069	0.072	0.140

Table 5.1: LMSE results of reflectance image estimates in different scenarios.

In Figure 5.4, we show one object of our dataset acquired with two different cameras and under two different illumination conditions (first column). In the second column we see the reflectance estimates provided by the method of Gehler *et al.* [60] when there is no available information about the light source or the camera and white light and sRGB sensors are assumed (SC4). In the last column, we show the reflectance estimates for the same method when photometric information is available (SC1). As expected, when the effects of the illuminant of the scene and the camera sensors are removed, the estimated reflectances are perceived as being more similar.

In this section we have provided ground truth data for intrinsic image evaluation which includes information about the illuminant of the scenes and the camera sensors. We have shown that any available knowledge about the color of the light source or the camera model can be used by existing methods in intrinsic image decomposition to provide considerably better reflectance estimates. We conclude that the general framework introduced in Chapter 4, which models the effects of both the color of the illuminant and the camera sensors, should be used in further approaches in intrinsic image decomposition.

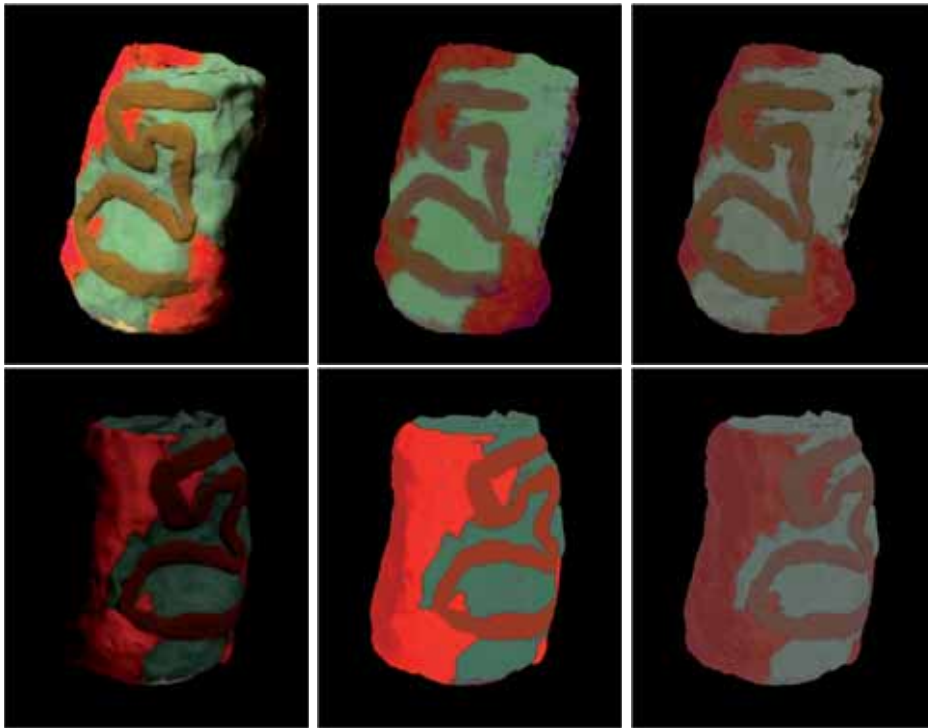


Figure 5.4: Removing the effect of the illuminant and the camera sensors increases the similarity between the different estimated images. In the upper row we see an object from the dataset acquired with the Nikon D5200 camera and under a reddish illuminant, while in the lower row we see the same object acquired with the Sigma Foveon D10 camera under a bluish illuminant. In the first column we see the original input images. In the second column we observe the reflectance estimates according to Gehler's method [60] when we assume white light and sRGB sensors. Finally, in the third column we see the reflectance estimates of the same method when information about the camera model and the illuminant of the scene is available and the effects of both factors have been removed.

## 5.2 Synthetic Intrinsic Image Dataset

In this section we present a synthetic dataset for intrinsic image evaluation which includes not only scenes displaying a single object illuminated by white light, but also scenes including multiple objects and multiple non-white illuminants with complex settings leading to interreflections. The objective of this new ground truth collection is to overcome the shortcomings of the existing datasets for 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<sup>†</sup> to further motivate research into more complex reflectance models.

Recently, advances in digital 3D modeling software have enabled users to rely on these rendering 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 model of the 3D scene 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 known illumination and reflectance models [126].

In this work, we have used Blender [1] to model the 3D scenes and YafaRay [5] to render 2D images from these scenes. Both of these applications are free and open source.

### 5.2.1 Motivation

In Chapter 4, we have seen that algorithms for intrinsic image decomposition can be distinguished by their assumptions on the underlying reflectance models. We will use Shafer’s reflection model [135], described in Chapter 4, to demonstrate the differences between existing datasets and our dataset. In the MIT dataset [71] the illuminant is considered to be white. This assumption is shared by most of the methods for intrinsic image estimation [20, 60, 133]. Recently, Barron and Malik [19] relaxed this assumption: they allowed the illuminant color to vary but only considered direct illumination (ignoring interreflections). They constructed a dataset by synthetically relighting the objects of the MIT dataset [19].

However, obtaining a precise ground truth for complex real scenes, such as landscapes, would be impracticable using the procedure described in [71]. Recently, the use of synthetic data to train and test complex computer vision tasks has attracted growing attention due to the increased accuracy of rendering engines to represent the world. In addition synthetic data allows for easy access to the ground truth. Marin *et al.* [116] and Vazquez *et al.* [150] showed that a pedestrian detector trained from virtual scenarios can obtain competitive results on real-world data. Liebelt and Schmid [107] used synthetic data to improve multi-view object class detection. Finally, Rodriguez *et al.* [128] generated synthetic license plates to train recognition system.

---

<sup>†</sup>[http://www.cic.uab.cat/Datasets/synthetic\\_intrinsic\\_image\\_dataset](http://www.cic.uab.cat/Datasets/synthetic_intrinsic_image_dataset)

In this work, we create a synthetic dataset by using rendering techniques from the field of computer graphics. This allows us to remove the restriction other datasets put on the illumination of the scene. In our dataset, 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 dataset of intrinsic images which considers these more complex reflections. Later in this section we analyze rendering accuracy for such reflection phenomena.

Note that Shafer’s reflection model assumes that the materials have Lambertian reflectances. Even though specular materials can be accurately rendered, we exclude them from this dataset because most existing algorithms of intrinsic image decomposition are not able to handle non-Lambertian materials.

### 5.2.2 Global Lighting for Scene Rendering

In order to obtain more photo-realistic lighting for 3D scene rendering, a group of rendering algorithms has been developed which is 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.

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, which was developed by Henrik Wann Jensen [88]. To achieve physically sound results and photo-realism in our dataset we make use of the photon mapping method embedded in YafaRay. Figure 5.5 shows the importance of indirect lighting in order to render more realistic images. For this purpose we compare the final renderings of our dataset (a) to the renderings which only consider direct lighting (images on the left). The diffuse interreflections found in the final renderings (b) contribute to provide a stronger sense of realism to these images.

### 5.2.3 Analysis of Color Rendering Accuracy

Synthetic datasets should accurately model the physical reality of the real world in order to be useful to train and evaluate computer vision algorithms. In this section we analyze the accuracy of color rendering based on the diagonal model, as it is typically done in the field of computer graphics.

Full multispectral data is computationally very expensive. For this reason, rendering engines approximate Equation 4.9 with

$$\hat{p}_k = \int_{\Lambda} f_r(\lambda) S_k(\lambda) d\lambda \int_{\Lambda} L(\lambda) S_k(\lambda) d\lambda, \quad (5.2)$$

where the integrals are over all wavelengths  $\lambda$  of the visible spectrum  $\Lambda$ . In vector



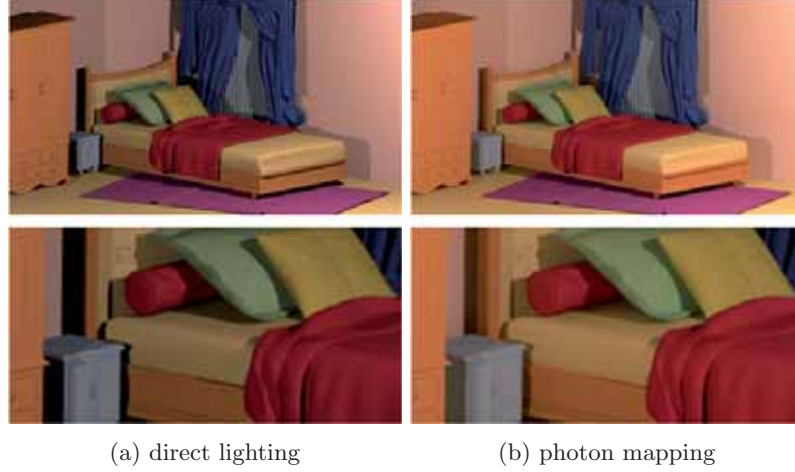


Figure 5.5: Comparing different rendering methods: *direct lighting* (a) and *photon mapping* (b).

notation we could write this as

$$\hat{\mathbf{p}} = \mathbf{f}_r \circ \mathbf{L}, \quad (5.3)$$

where we use bold symbols to denote vectors (*i.e.*  $\hat{\mathbf{p}} = [\hat{p}_1, \dots, \hat{p}_n]$ , where  $n$  is the number of sensors  $k$  considered),  $\circ$  is the Hadamard product, and

$$\mathbf{f}_r = \int_{\Lambda} f_r(\lambda) S_k(\lambda) d\lambda \quad (5.4)$$

and

$$\mathbf{L} = \int_{\Lambda} L(\lambda) S_k(\lambda) d\lambda. \quad (5.5)$$

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 lighting and interreflection from another surface we can write:

$$\hat{\mathbf{p}} = \mathbf{f}_r^1 \circ \mathbf{L} + \mathbf{f}_r^2 \circ \mathbf{f}_r^1 \circ \mathbf{L}, \quad (5.6)$$

where the superscript is used to distinguish the material reflectance of different objects. The accuracy of the approximations in Equation 5.3 and Equation 5.6 depends on the shape and the number of sensors  $k$  considered. Typically rendering machines apply three sensors, however it is known that rendering accuracy increases when considering more sensors [44, 40]. These additional sensors preserve additional spectral information, which results in better representations of complex lighting effects such as interreflections.

To test the accuracy of the approximate values  $\hat{p}_k$  we perform a statistical analysis. We use the 1269 Munsell color patches [4] and we compute both  $p_k$  and  $\hat{p}_k$  (using

Equations 4.9 and 5.2, respectively). For sensors  $S_k$  we use Gaussian shaped sensors which are equally spaced over the visible spectrum. We compare the reconstruction error

$$\varepsilon = \frac{\|\mathbf{p}(\mathbf{x}) - \hat{\mathbf{p}}(\mathbf{x})\|_2}{\|\mathbf{p}(\mathbf{x})\|_2}. \quad (5.7)$$

for the cases of three, six and nine sensors. We consider both reflections with a single bounce (Equation 5.3) and with two bounces (Equation 5.6). We use the standard D65 daylight illuminant. Dark patches were discarded because they cause the reconstruction error to be unstable.

sensors	One bounce			Two bounces		
	mean (%)	s.d. (%)	max (%)	mean (%)	s.d. (%)	max (%)
3	0.58	0.44	2.88	1.38	1.62	23.84
6	0.19	0.14	1.25	0.55	0.55	9.06
9	0.12	0.08	0.86	0.34	0.29	3.77

Table 5.2: Reconstruction error for single and two bounce reflection for 3, 6, and 9 sensors.

Table 5.2 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%. Observe that both the mean and standard deviation of the errors decreases considerably when six or nine sensors are being used. 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 [44, 40]. The resulting 6 channel image is projected back to a RGB image using linear regression. Although our results show that using 9 sensors would provide more accurate renderings, we have considered that the computational cost of using more than 6 sensors is too high. In the only dataset of intrinsic images containing multi-illuminants [19], illuminants were introduced synthetically by using a 3 channel approximation. 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 using the ratio

$$r = \frac{\sum_{\mathbf{x}} \|\mathbf{p}^1(\mathbf{x})\|}{\sum_{\mathbf{x}} \|\mathbf{p}^\infty(\mathbf{x})\|} \quad (5.8)$$

where  $\mathbf{p}^\infty$  is the final rendering and  $\mathbf{p}^1$  is the single bounce rendering. For the 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, thus providing a stronger sense of realism to our rendered images.

#### 5.2.4 Proposed dataset

Our dataset consists of two sets of images: single objects and complex scenes. Our aim is to use the first set of images to simulate scenes which are similar to the scenes of the 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 we know their multispectral reflectance values. 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 [17].



Figure 5.6: Single object scenes included in our dataset.

Figures 5.6 and 5.7 show, respectively, the single object and complex scenes included in our dataset. Figure 5.8 shows examples of the ground truth we provide with the dataset. Finally, Figure 5.9 shows two scenes of the dataset under different illumination conditions.



Figure 5.7: Complex scenes included in our dataset.



Figure 5.8: Two examples of ground-truth decomposition. For both rows, from left to right: the rendered scene, reflectance component, and shading-illumination.

### 5.2.5 Experiments

In order to show that our dataset is suitable to evaluate methods for intrinsic image decomposition, we compare four different methods for intrinsic image decomposition which were state-of-the-art [19, 60, 133, 90] at the time of publication. For this experiment, we have used the methods whose codes were publicly available, keeping 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 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.

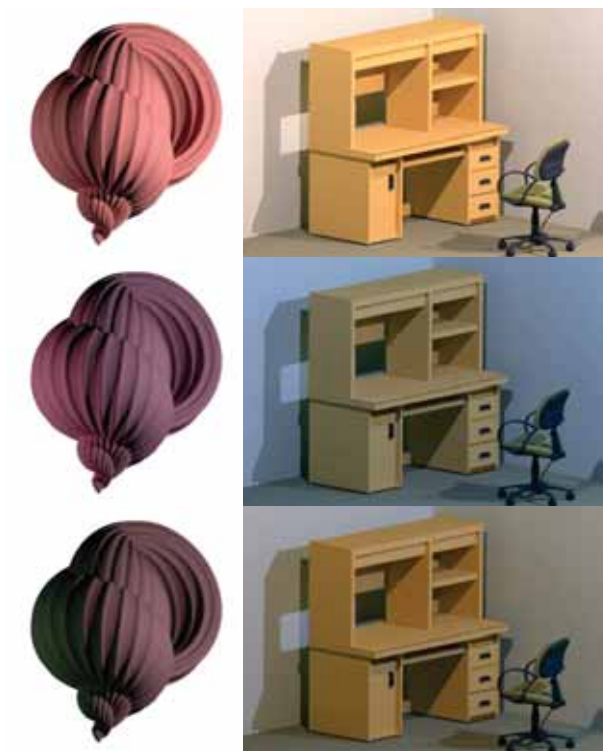


Figure 5.9: Two examples from the dataset under different illumination conditions. For both columns, from top to bottom: white illuminant, single-colored light, and two distinct colored illuminants.

Errors have been evaluated with the local mean squared error (LMSE) and considering the three RGB channels of the color image [71]. As reflectance images can be recovered only up to a scale factor, we have multiplied the estimated reflectance images by an  $\alpha$  factor which has been fitted for each local patch to minimize the MSE.

Tables 5.3 and 5.4 summarize the results for the reflectance and shading estimates of different approaches for intrinsic image decomposition. The error for all methods varies accordingly to the complexity of the illumination conditions, being lower when scenes are illuminated by white light and increasing when multiple light sources and colored illuminants are considered. 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.* [133] 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, since visual ranking of the accuracy did often not agree with the LMSE error based ranking. Observe that the method of Barron and Malik [19] needs masked input images, since it uses a prior on shape based on contours. Therefore, we just applied this method to the complex scenes in our dataset for the sake of completion, defining masks of the same size of the images.

Method	Single Objects					
	Reflectance			Shading		
	WL	1L	2L	WL	1L	2L
Barron & Malik [19]	0.082	0.099	0.102	0.043	0.046	0.054
Gehler <i>et al.</i> [60]	0.089	0.113	0.123	0.043	0.045	0.051
Serra <i>et al.</i> [133]	0.063	0.069	0.076	0.021	0.022	0.025
Jiang <i>et al.</i> [90]	0.160	0.161	0.161	0.040	0.043	0.050

Table 5.3: LMSE results of three intrinsic image methods on the single object scenes included in our dataset. Errors for reflectance and shading are given separately for the sake of clarity. Results for white illumination (WL), one illuminant (1L), and two illuminants (2L) are averaged.

Method	Complex Scenes					
	Reflectance			Shading		
	WL	1L	2L	WL	1L	2L
Barron & Malik [19]	0.020	0.059	0.039	0.011	0.014	0.014
Gehler <i>et al.</i> [60]	0.018	0.067	0.040	0.007	0.009	0.009
Serra <i>et al.</i> [133]	0.027	0.041	0.033	0.006	0.006	0.007
Jiang <i>et al.</i> [90]	0.069	0.070	0.070	0.013	0.013	0.013

Table 5.4: LMSE results of three intrinsic image methods on the complex scenes included in our dataset. Errors for reflectance and shading are given separately for the sake of clarity. Results for white illumination (WL), one illuminant (1L), and two illuminants (2L) are averaged.

## 5.3 Discussion and Conclusion

In this chapter, we have presented two new datasets for intrinsic image evaluation.

First, we focused on the importance of the effects of both the illuminant of the scene and the camera sensors during the image acquisition process (described in Chapter 4). In order to demonstrate how these factors affect the estimates of existing methods of intrinsic image decomposition, we built a calibrated dataset which includes ground truth information on the effects of the illuminant and the camera sensors. This dataset is one sound contribution of this thesis.

In order to validate the general formulation that we introduced in Chapter 4, we used our real dataset to evaluate different methods of intrinsic image estimation. Our results indicate the importance of modeling the effects of the color of the illuminant and the camera sensors in the formulation of intrinsic image decomposition in order to obtain more accurate reflectance estimates.

Second, we encouraged the collection of large intrinsic synthetic image datasets. These datasets allow the evaluation of methods of intrinsic image decomposition in complex scenes and under complex illumination conditions. In order to do that, we presented a synthetic dataset containing both simple objects, resembling those in the MIT dataset [71], and complex scenes. All the scenes in the dataset have been rendered under different illumination conditions, ranging from a single white illuminant to multiple colored light sources. In this dataset, complex illumination effects such as interreflections have also been modeled. This dataset is another important contribution of this work.

We validated our dataset with different methods, and showed that the results we got are consistent with the ones these methods achieved when they were evaluated with the MIT dataset. Therefore, we claim that it is possible to apply the idea of knowledge transfer that has been successfully used in other fields such as pedestrian detection [116, 150] to the problem of intrinsic image estimation.

The use of synthetic data enables the collection of large and diverse datasets for intrinsic image evaluation, and rendering engines provide straightforward and accurate ground truth data for multiple scene characteristics.

Finally, we include our datasets to Table 2.5, presented in Chapter 2, in order to compare our ground truth collections with the other datasets that have been used so far for intrinsic image evaluation.

Dataset	Acquisition		Ground Truth				
	Data	Objects	Shad. & Refl.	Specularities	Depth	Illuminant	Sensors
Crayola [145]	Laboratory	46	✓				
MIT [71]	Laboratory	20	✓	✓			
extended MIT [20]	Synthetic	20	✓		✓		✓
MPI-Sintel [35]	Synthetic	35*	✓		✓		
Bell <i>et al.</i> [28]	Natural	5000	✓				
Our Synthetic Dataset [24]	Synthetic	17	✓				✓
Our Calibrated Dataset	Laboratory	20	✓				✓

Table 5.5: Comparison of the existing datasets for intrinsic image evaluation. \* Sequences with different numbers of frames (50 on average).



# Chapter 6

## Conclusions and Future Work

In this thesis we have analyzed many aspects of one fundamental problem in the field of computer vision: the decomposition of a single image into its intrinsic characteristics. We have also presented several novel contributions to this problem. From these contributions we have drawn multiple conclusions, which have already been exposed separately in the previous chapters. All these conclusions are reviewed together in this chapter. We also suggest some possible extensions and applications of the research lines that have been described in this thesis and discuss the recent trends in the field of intrinsic images and their potential developments.

### 6.1 Conclusions

We introduced the problem of intrinsic image estimation in the field of computer vision in Chapter 1. We also presented the background of the problem, showing the relationship of intrinsic images with art techniques, the human visual system and human perception. Additionally, we mentioned topics such as color constancy or sensor calibration whose photometric effects have a big influence in the problem of intrinsic image estimation but had never been considered in the field so far.

In Chapter 2 we provided an exhaustive overview on the different existing methods for intrinsic image estimation, taking into account different information such as the assumptions that authors make about the world or the visual cues the methods use to constrain the space of solutions of the decomposition problem. We also presented the multiple datasets and metrics that have been used so far to evaluate these methods. This review of the intrinsic image literature in the field of computer vision is one of the contributions of this thesis.

We observed that emerging methods are using input information from multiple sources such as video, image collections or RGB-D images. This extra information simplifies the problem of intrinsic image decomposition. At the same time, deep learning methods and advanced optimization techniques such as high-order MRFs are being used in order to jointly estimate multiple intrinsic characteristics of the scene, unifying different problems such as shape-from-shading, color constancy or shadow-removal.

The joint estimation of multiple intrinsic characteristics may seem to add complexity to this already underconstrained problem. However, the inclusion of multiple intrinsic characteristics of the scene into the formulation of the problem allows the simplification of the decomposition problem, since most of these characteristics are closely related to each other, and each of them presents some regularities which can be exploited to bound the space of plausible solutions.

In Chapter 3 we presented a method for intrinsic image decomposition which combines observations from two color cues in a conditional random field. One of the cues is based on the semantic terms that human beings use to describe color. We showed that color-name descriptors, based on psychophysical data, provide a sparse set of color values to describe the reflectances in the scene. Color sparsity makes color names more robust for describing object reflectance than any color descriptor based on standard color spaces, such as RGB values.

The second cue is based on an analysis of color distributions in the histogram space and provides a consistent description of surfaces sharing the same reflectance. This attribute captures the continuity of material color through shading variations, enhancing the stability of color names against strong changes in illumination due to shadows and highlights.

In our method, a graph cut algorithm is used to minimize the energy function, and a post-processing step is applied in order to enforce the smoothness of the shading by locally modifying the intensity of the different reflectance descriptors in the reflectance intrinsic image. The estimation of intrinsic reflectance images using a combination of these color cues is a sound contribution of this thesis [133].

In Chapter 4 we described a generic theoretical framework for intrinsic image estimation. The proposed mathematical formulation includes information about the color of the illuminant and the effects of the camera sensors, both of which contribute to the observed color of the reflectance of the objects in the scene during the acquisition process. This model allows us to represent a wide range of intrinsic image decompositions depending on the specific assumptions on the geometric properties of the scene configuration and the spectral properties of the light source and the acquisition system, thus unifying previous models in a single general framework. By modeling these effects, we get a “truly intrinsic” reflectance image, called absolute reflectance, which is invariant to changes of illuminant or camera sensors.

We validated our general intrinsic image framework experimentally with both synthetic data and natural images, and showed that the effects of both the illuminant of the scene and the camera sensors are critical for the problem of intrinsic image decomposition. Moreover, we demonstrated that even partial information about the illuminant and the camera sensors improves significantly the estimated reflectance images, thus making our method widely applicable.

The definition and validation of this theoretical framework for intrinsic image decomposition and the definition of absolute reflectances are important contributions of this thesis [134].

In Chapter 5 we presented a dataset for intrinsic image evaluation which includes ground truth data about the illuminant of the scene as well as the camera sensors. This dataset was also used to validate our theoretical framework, and is in its own another contribution of this thesis.

From our experiments we derived that existing methods which assume white light and sRGB sensors carry big errors in their intrinsic reflectance estimates. We concluded that it is necessary that future methods for intrinsic image estimation model the effects of both the illuminant of the scene and the camera sensors.

Finally, we presented a synthetic dataset containing both single objects and complex scenes under multiple illumination conditions. We analyzed the accuracy of color rendering based on the diagonal model and showed that the use of 6 sensors, instead of 3 (RGB sensors are commonly used), resulted in better rendered images. We also showed that current methods can be evaluated with such a dataset, and the results they provide with our dataset are coherent with those achieved with the MIT dataset.

We concluded that computer graphics software and rendering engines allow us to easily build large synthetic datasets. Moreover, ground truth information for multiple intrinsic characteristics can be directly computed. Therefore, synthetic ground truth collections offer a realistic alternative to existing datasets for intrinsic image evaluation. This dataset is another contribution of this thesis [24].

To sum up, in this thesis we have studied the problem of intrinsic images in the field of computer vision from multiple perspectives. To start with, we have proposed a method for intrinsic image estimation. We have next defined a theoretical model for intrinsic image decomposition. We have built two new datasets for intrinsic image evaluation, acquiring ground truth data from natural images in a lab and rendering synthetic scenes using computer graphics software and rendering engines. The principal novel contributions of this work are listed below.

- This thesis provides an exhaustive review of the intrinsic image literature in the field of computer vision.
- This thesis uses color cues based on psychophysical data and color distributions for the estimation of intrinsic reflectance images.
- This thesis defines and validates a theoretical framework for intrinsic image decomposition which includes information about the color of the illuminant and the effects of the camera sensors and extends previous existing formulations.
- This thesis also defines a new intrinsic term, called absolute reflectance, which is invariant to changes of illuminant or camera sensors.
- This thesis presents a new dataset for intrinsic image evaluation which contains information about the illuminant of the scene and the camera sensors.
- This thesis introduces a synthetic collection of ground truth data for intrinsic image evaluation containing both simple objects and complex scene under different illumination conditions.

## 6.2 Future Work

This thesis provides multiple novel contributions in the field of intrinsic image estimation. These contributions can be considered as the starting point for new lines of

research. In this section we will first present some of our ideas to extend and improve our work, and then will provide a general analysis on the recent trends in intrinsic image decomposition.

### 6.2.1 Future Research and Applications

We present here some proposals to extend the lines of research presented in this thesis as well as some possible applications of our contributions.

We have multiple ideas in order to improve the method for intrinsic image estimation that we presented in Chapter 3. In this work, we discretized the set of possible labels by allowing combinations of at most 3 different labels taking 4 different non-zero values and imposing their sum to be 1. Increasing the number of labels does not seem to make any sense since in the color-name descriptor used in our method [30] there are just a few values located in junctions of more than 3 different colors. However, increasing the number of possible non-zero values, could lead to a more robust representation of the reflectances in the image. Another idea is the definition a better metric, based on human perception, to measure the distance between two labels. Although we tested, in this thesis, different confusion matrices based in the psychophysical data from Benavente *et al.* work [30], the  $\alpha$ -expansion graph cut algorithm that we used to optimize our energy function forced us to define a metric satisfying the triangle rule. A final idea in this direction is the definition of a new set of weights in the pair-wise potential, based on observations of the color shade descriptor, which could lead to better reflectance estimates.

In a different direction, we also propose to extend our method to jointly estimate illumination and shape properties of the scene, similarly to what Barron and Malik did in [19]. This could be done by using high-order MRFs and redefining our energy function accordingly using existing methods on shape-from-shading and color constancy which are also defined in probabilistic frameworks. Using RGB-D images as an input for our method would also be interesting since our method focuses on estimating the reflectance image, and an approximate depth information input could be used to impose some restrictions on the shading image, thus improving the inference on the reflectance component.

Balagué explored some of the ideas mentioned above in [15]. The author analyzed the influence of the number of possible color-name labels on the results. He increased the number of possible non-zero values and showed that when the number of possible labels was bigger, the estimated color values in the reflectance images looked qualitatively closer to the colors in the ground truth reflectance images. However, the quantitative results did not improve significantly and computational time was dramatically increased. Balagué also performed a thorough analysis of different confusion matrices by using an optimization algorithm which was less restrictive. Nonetheless, the use of different confusion matrices did not result in better intrinsic estimates. Finally, Balagué also tried to modify the structure of the MRF by defining a new set of weights in the pair-wise potential. Adding these new weights did not result in any significant improvement. However, the author suggested that defining a continuous function to determine these weights, instead of using thresholds, could result in better intrinsic image estimates. Although Balagué presented in [15] a first analysis on some

of our ideas, a thorough analysis still needs to be done.

In Chapter 4, we concluded that the effects of the illuminant of the scene and the camera sensors influenced a lot the performance of existing methods in intrinsic image estimation. Although some methods have already considered the illuminant of the scene [19, 21, 38] in their formulation, to the best of our knowledge camera sensors have been completely ignored so far. These photometric effects should be modeled in future approaches for intrinsic image decomposition.

In big collections of images, such as Flickr, information about the camera device used to acquire the image is usually available. We propose to use this information to build a method that, given an image, estimates the model of the camera that was used to take the image, in a similar fashion to what we did in Chapter 4, but using more images and cameras. Such a camera estimator would simplify the problem of recovering absolute reflectance images and could also be applied to other fields such as image forensics, which focuses on identifying the source of a digital images without having any prior information about these images.

Multi-view stereo techniques infer the shape of a scene given multiple pictures of this scene taken from different points of view. If we were able to remove illuminant and camera sensor effects from images, this would benefit multi-view stereo techniques, since the big collections of images that exist on the internet could be used to recover the shape of multiple scenarios. For example, we could use multiple images of the Sagrada Família from Flickr, taken with different cameras and under different illumination conditions, to recover the 3D shape of the sanctuary.

## 6.2.2 Recent Trends in the Field of Intrinsic Images

After reviewing the literature on intrinsic images in computer vision in Chapter 2, we observed that the recent trends in the field are focused on the joint estimation of multiple intrinsic components, including scene characteristics such as the depth and orientation of the objects or the color and direction of the illuminants in the scene. This is an important step towards the unification of multiple problems in the field of computer vision which estimate different intrinsic characteristics of the scene, such as shape-from-shading, color constancy or specular removal techniques.

Other problems related to high-level attributes of a scene such as object segmentation, semantic labeling or optical flow, have also been combined with the problem of intrinsic image estimation achieving great results.

In our opinion, the future of this joint estimation will be basically affected by two factors. The first one is the use of extra input information which may simplify the decomposition problem, such as RGB-D images, videos or internet image collections. This factor is closely related to the development of technology.

The other is the development of powerful computational techniques able to optimize complex energy functions such as those defined with high-order MRFs, fast and efficiently. Other advanced techniques, such as deep learning methods, have recently provided promising results in the field of computer vision as well.

The evaluation of the problem must evolve in conjunction with the methods. Since most current methods estimate multiple intrinsic components, we emphasize the urge

to build larger datasets which contain accurate ground truth data of these components, thus enabling the quantitative evaluation of actual and future approaches.

Since building datasets for intrinsic image evaluation has proved to be a challenging problem, we propose the use of synthetic data to build future datasets. Nowadays, extremely realistic scenes can easily be reproduced using computer graphics software and rendering engines. Moreover, exact information from multiple intrinsic components of the scene can be recovered straightforwardly.

Although synthetic datasets have already been used for intrinsic image evaluation, a lot of work still needs to be done in this direction. The MPI-Sintel dataset [35] was not specifically created for the purpose of intrinsic image evaluation, and the new synthetic dataset [24], presented in Chapter 5, has just been a first step towards the use of synthetic data for intrinsic image evaluation. However, new collections of synthetic data, containing a larger number of complex scenes and including ground truth information about multiple intrinsic components, are still necessary.

We also observed the drawbacks of the multiple existing measures for intrinsic image evaluation. It is important to define new standard metrics which overcome the problems of previous similarity measures. In our opinion, each intrinsic component should be evaluated with a different metric. In particular, special attention should be given to the evaluation of color differences in the reflectance image, since Euclidean distances in most color spaces are not correlated with human perception.

# Appendix A

## Discretization of the color-name descriptor

Given an image, the output of the color-name descriptor consists of 11-dimensional arrays containing in each position the probability of a given RGB value to belong to each of the 11 universal color classes defined by Berlin and Kay in their anthropological work [31]. These vectors satisfy the two following properties:

- Their values are non-negative real numbers.
- Their values sum to 1.

In order to make our algorithm computationally efficient, we need to reduce the number of possible labels while preserving the two properties mentioned above. Thus, we discretize the set of possible values, making the non-zero values (three at most) to lay within the set  $\{0.25, 0.5, 0.75, 1\}$  and imposing their sum to be 1 .

This leads to four kinds of labels: a single value is non-zero, two values are non-zero (two cases), and three values are non-zero. Each of these kinds of label can be seen as a permutation with repetition of  $n$  elements, where the first element is repeated  $a$  times, the second  $b$  times, the third  $c$  times, etc. Therefore, we can calculate the number of possible labels of each kind using the following formula:

$$P_n^{a,b,c,\dots} = \frac{n!}{a!b!c!\dots}, n = a + b + c + \dots \quad (\text{A.1})$$

These are the four kinds of labels we can find:

- **Labels containing a single non-zero value (1):** Permutations of eleven elements, where the first element, 0, is repeated 10 times, and the second element, 1, just appears once.

$$P_{11}^{10,1} = \frac{11!}{10!1!} = 11$$

- **Labels containing two equal non-zero values (0.5, 0.5):** Permutations of eleven elements, where the first element, 0, is repeated 9 times, and the second

element, 0.5, twice.

$$P_{11}^{9,2} = \frac{11!}{9!2!} = 55$$

- **Labels containing two different non-zero values (0.25, 0.75):** Permutations of eleven elements, where the first element, 0, is repeated 9 times, and the second (0.25) and third (0.75) elements, once.

$$P_{11}^{9,1,1} = \frac{11!}{9!1!1!} = 110$$

- **Labels containing three non-zero values (0.5, 0.25, 0.25):** Permutations of eleven elements, where the first element, 0, is repeated 8 times, the second element, 0.5, once, and the third element, 0.25, twice.

$$P_{11}^{8,2,1} = \frac{11!}{8!2!1!} = 495$$

Although these theoretical computations give 671 possible labels, most of them are never found in practice. This makes sense because in the color-naming model we use [30] almost all color borders are shared by 3 colors or less. Therefore, many labels, such as the one defining a color as having probability 1/2 of being red and 1/2 of being green, never happen. Thus, only considering labels with up to three positive coordinates is enough to accurately describe the whole RGB space. In the end, only 250 different labels are actually used.



# Appendix B

## Illumination Conditions

A Planckian light source represents the radiation emitted by a black body at temperature  $T$ . Its spectral distribution is described by Planck's law, which is given by

$$R(\lambda, T) = \frac{c_1}{\lambda^5} \left( \exp \left\{ \frac{c_2}{T\lambda} \right\} - 1 \right)^{-1}, \quad (\text{B.1})$$

where the radiation,  $R$ , is a function of the wavelength,  $\lambda$ , and the temperature,  $T$ ; and  $c_1$  and  $c_2$  are constants. The sun and incandescent light sources are examples of Planckian light sources. Common examples of non-Planckian illuminants are light-emitting diodes (leds) and fluorescent lamps. The Planckian illuminants used in our experiments were selected to cover a big range of common color temperatures (Figure B.1(a)), while the non-Planckian light sources were randomly selected from Barnard *et al.* dataset [17] (Figure B.1(b)).

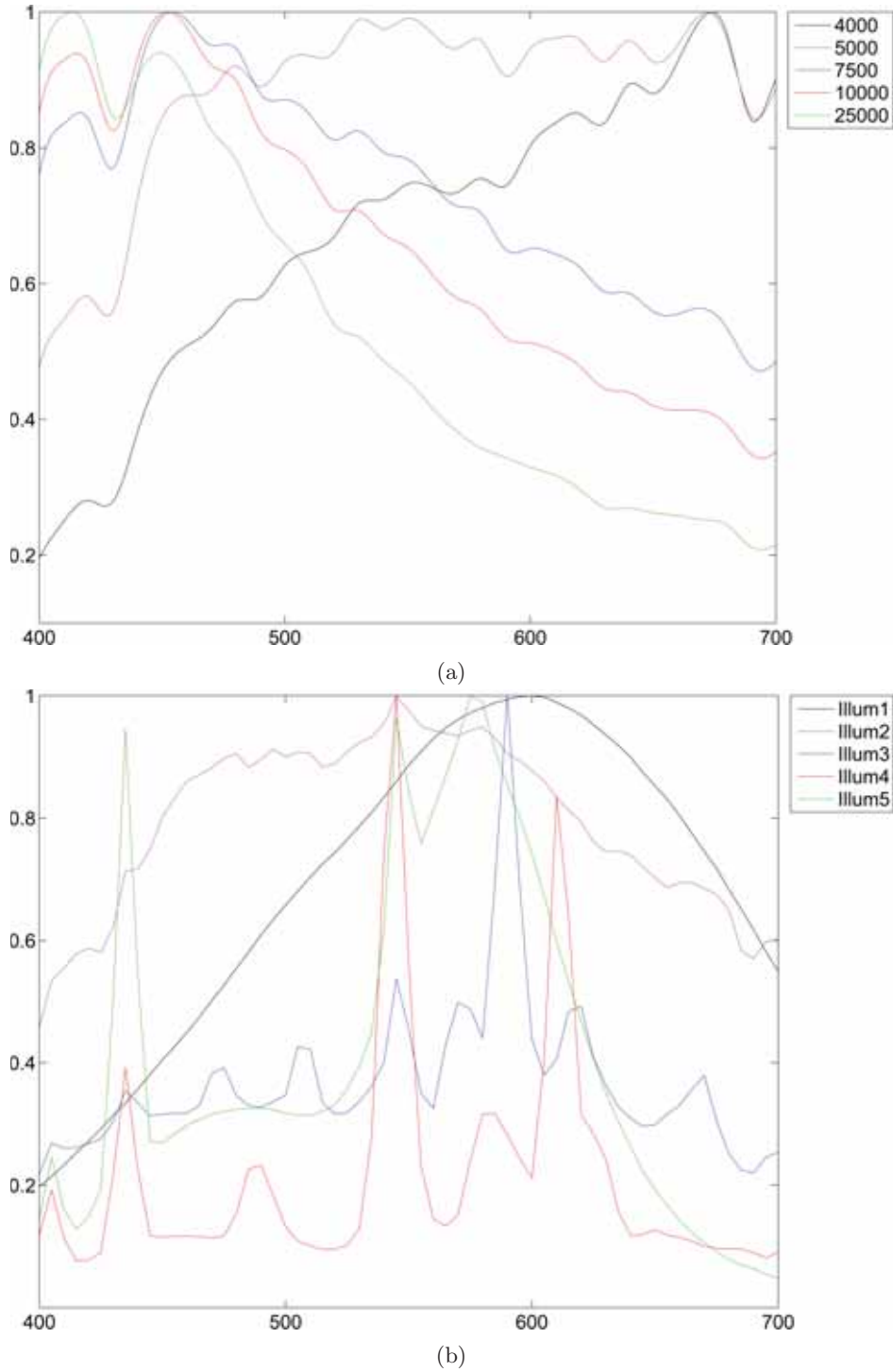


Figure B.1: Spectral power distribution of the 10 illuminants used in the experiment with synthetic data. (a) Planckian illuminants. (b) Non-Planckian illuminants.

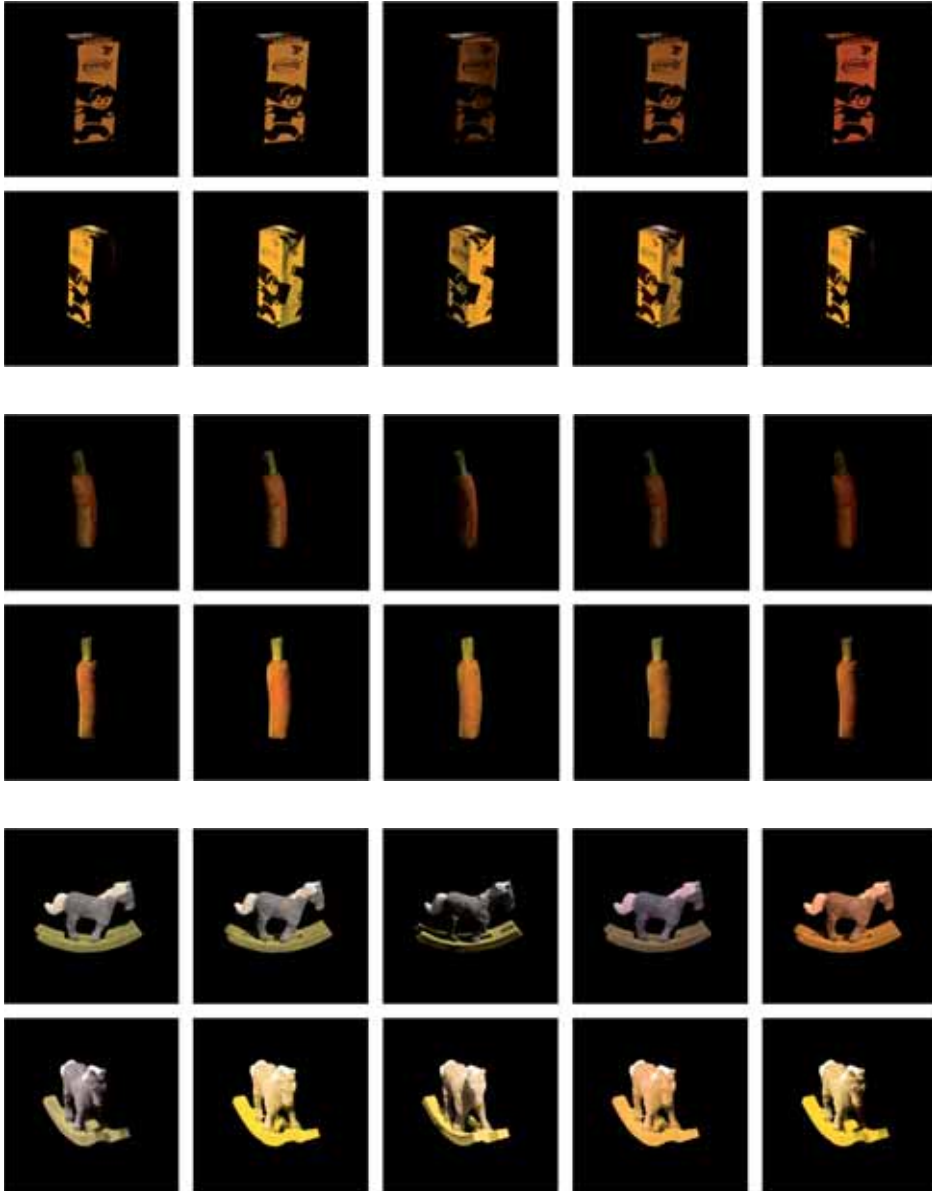
# Appendix C

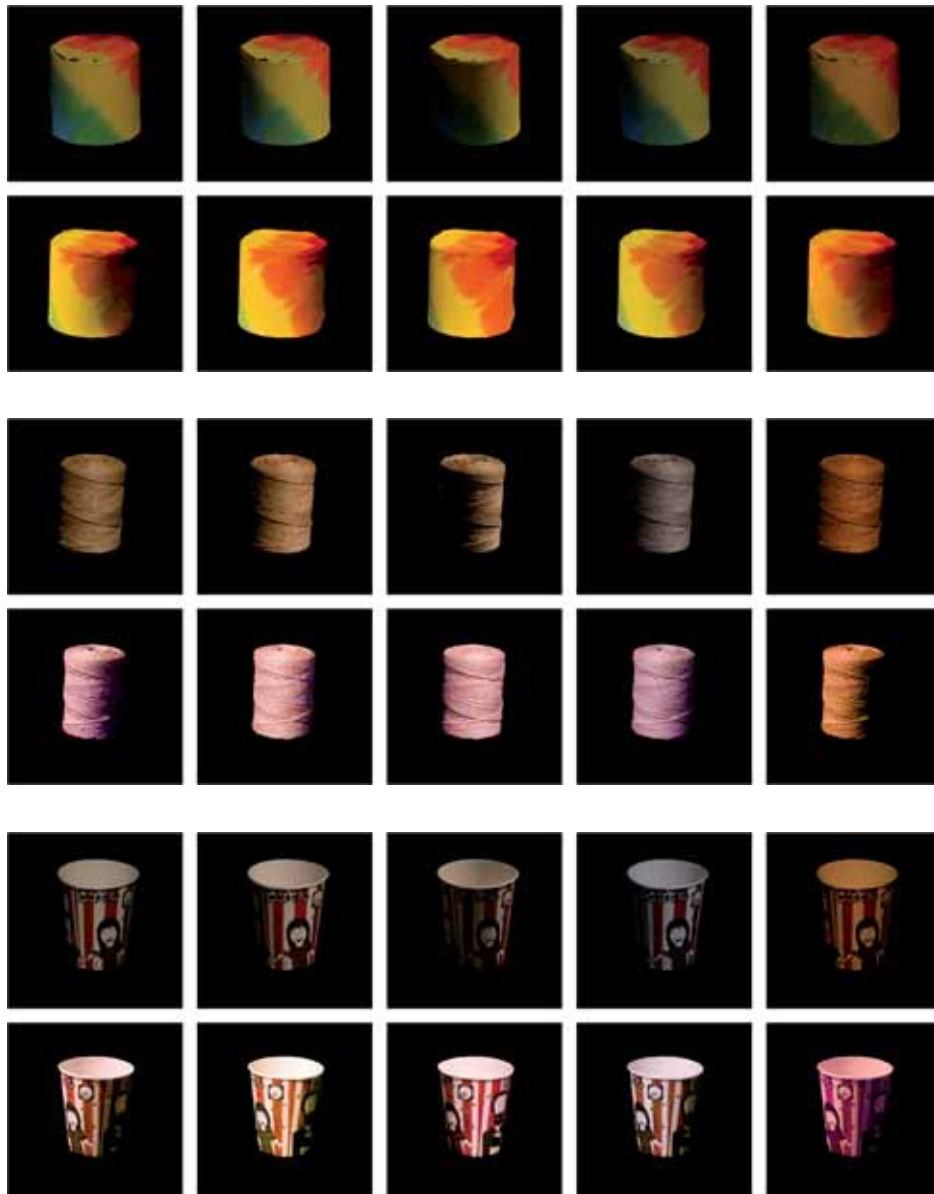
## Database images

We present here all the images included in the two datasets that we have already presented in Chapter 5.

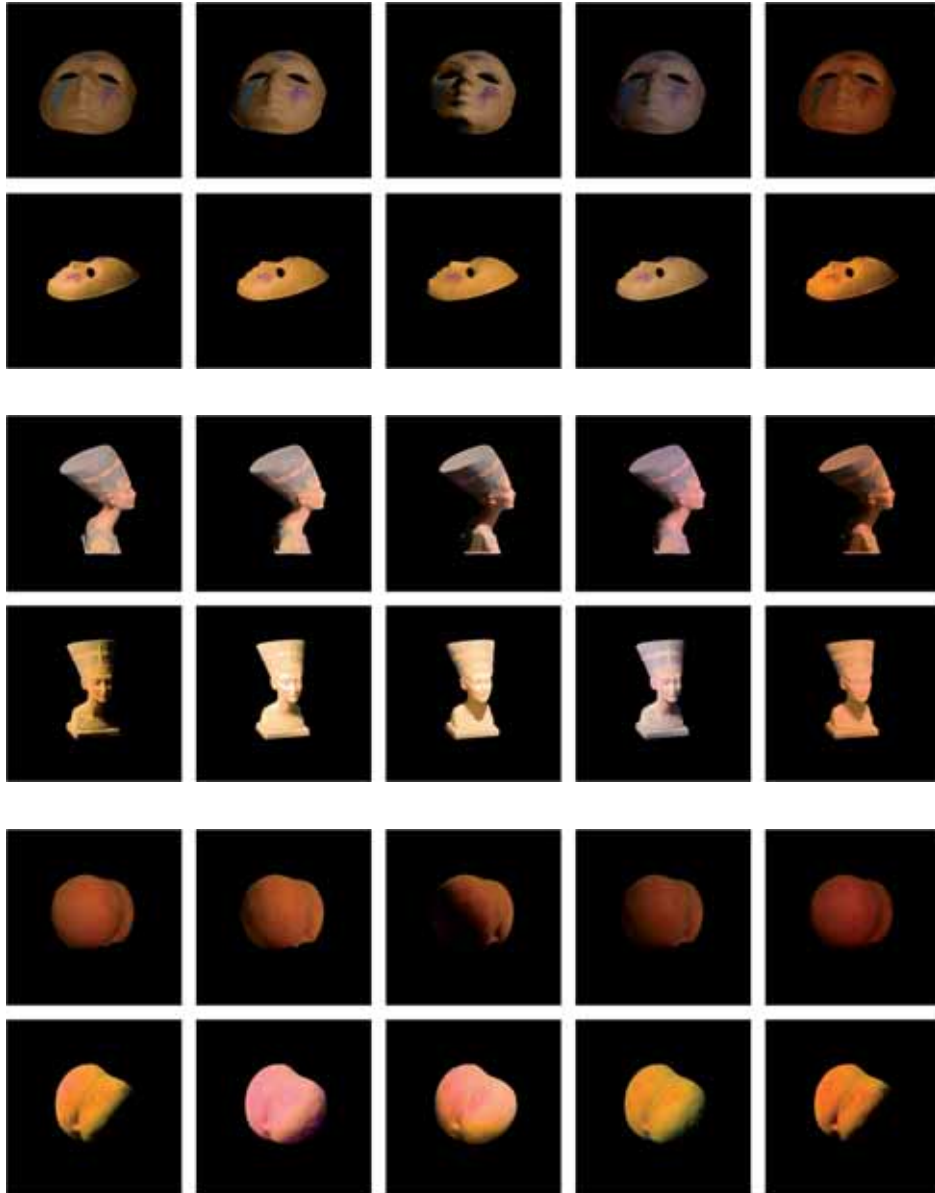
### C.1 Calibrated Dataset

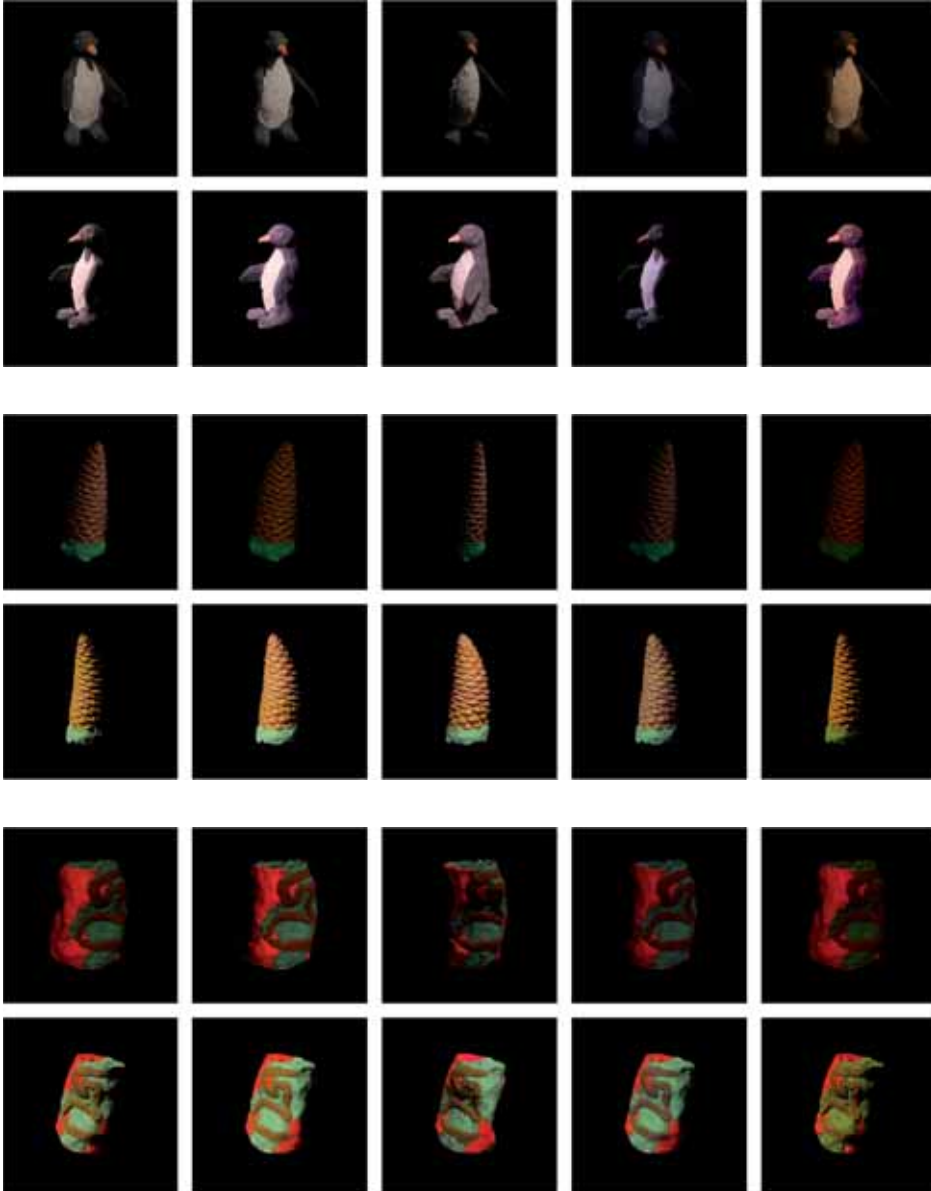
We show in this section all the objects in our calibrated dataset for intrinsic image evaluation described in Chapter 5. Our dataset contains 20 objects. Images of these objects have been acquired in a laboratory using 2 cameras, the Nikon D5200 and the Sigma Foveon D10, and 5 illumination conditions (a total of 200 images). All the objects are shown below. For each object, we see the images acquired with the Nikon D5200 camera in the first row, and the images acquired with the Sigma Foveon camera in the second row. Moreover, each column corresponds to one illumination condition. The first three columns correspond to the illumination conditions where no color filter have been applied. Thus, only the direction of the light source differs in these images. The images in the fourth column have been acquired under a bluish illuminant (*i.e.* a blue color filter has been placed in front of one of the light bulbs). The last column corresponds to the illumination condition where a yellowish color filter has been placed in front of one of the light bulbs.



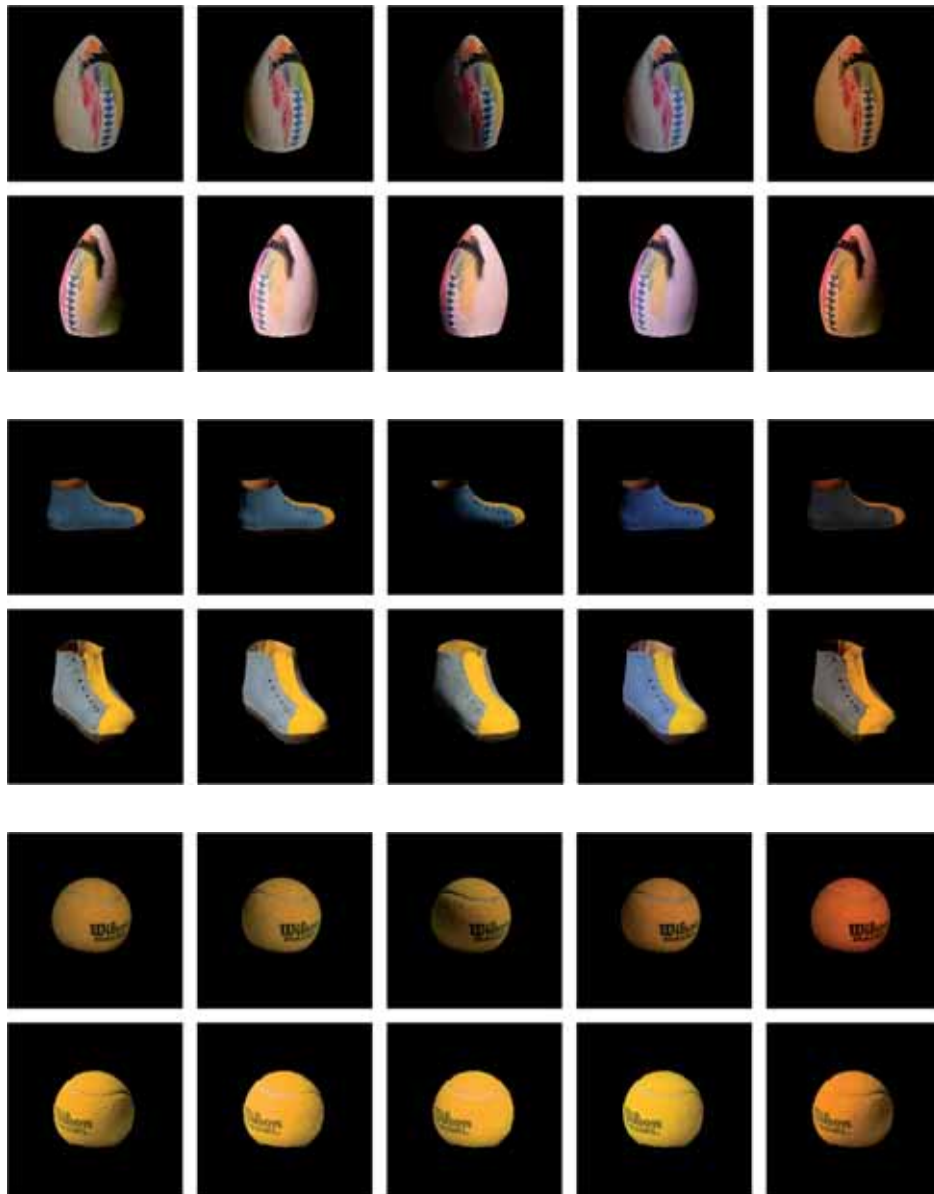


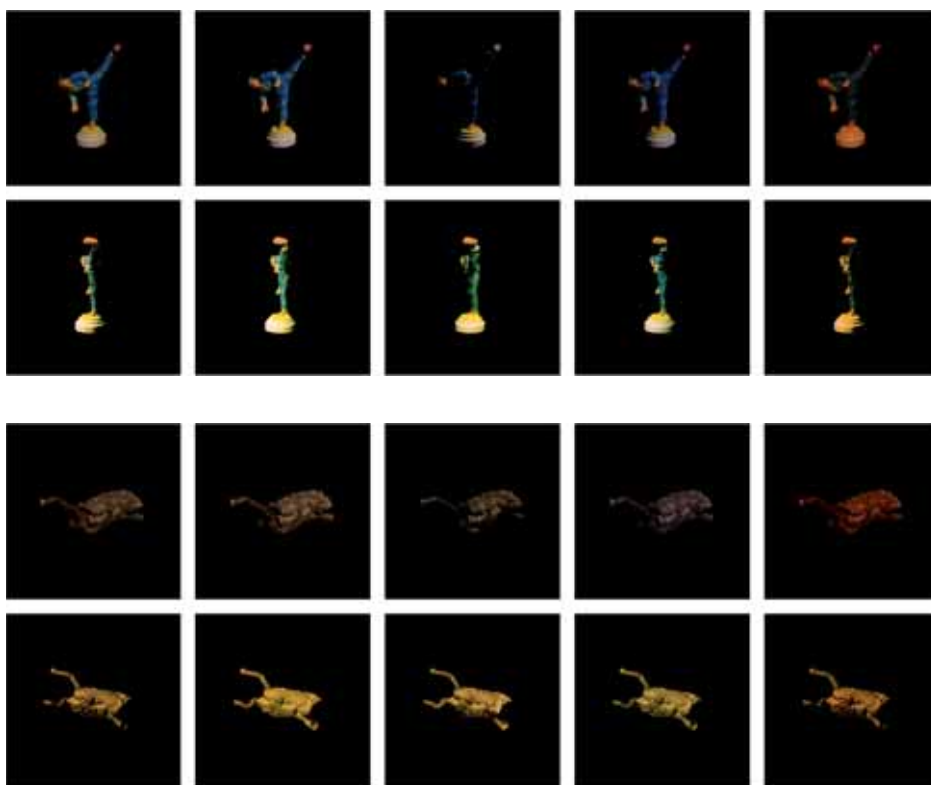












## C.2 Synthetic Dataset

We show in this section all the objects in our synthetic dataset for intrinsic image evaluation described in Chapter 5. Our dataset contains 68 images. These images represent both simple objects and complex scenes. In Figure C.1 we show the 8 scenes with a single object, which have been rendered under four different illumination conditions (32 images of single objects). In Figure C.1 we show three different points of view of 3 complex scenes under four different illumination conditions (a total of 36 images representing complex scenes). In both figures, the images in the first column are illuminated with white light, the images in the second column contain one colored light, and the images in the third and fourth columns are rendered with two illuminants with distinct colors.

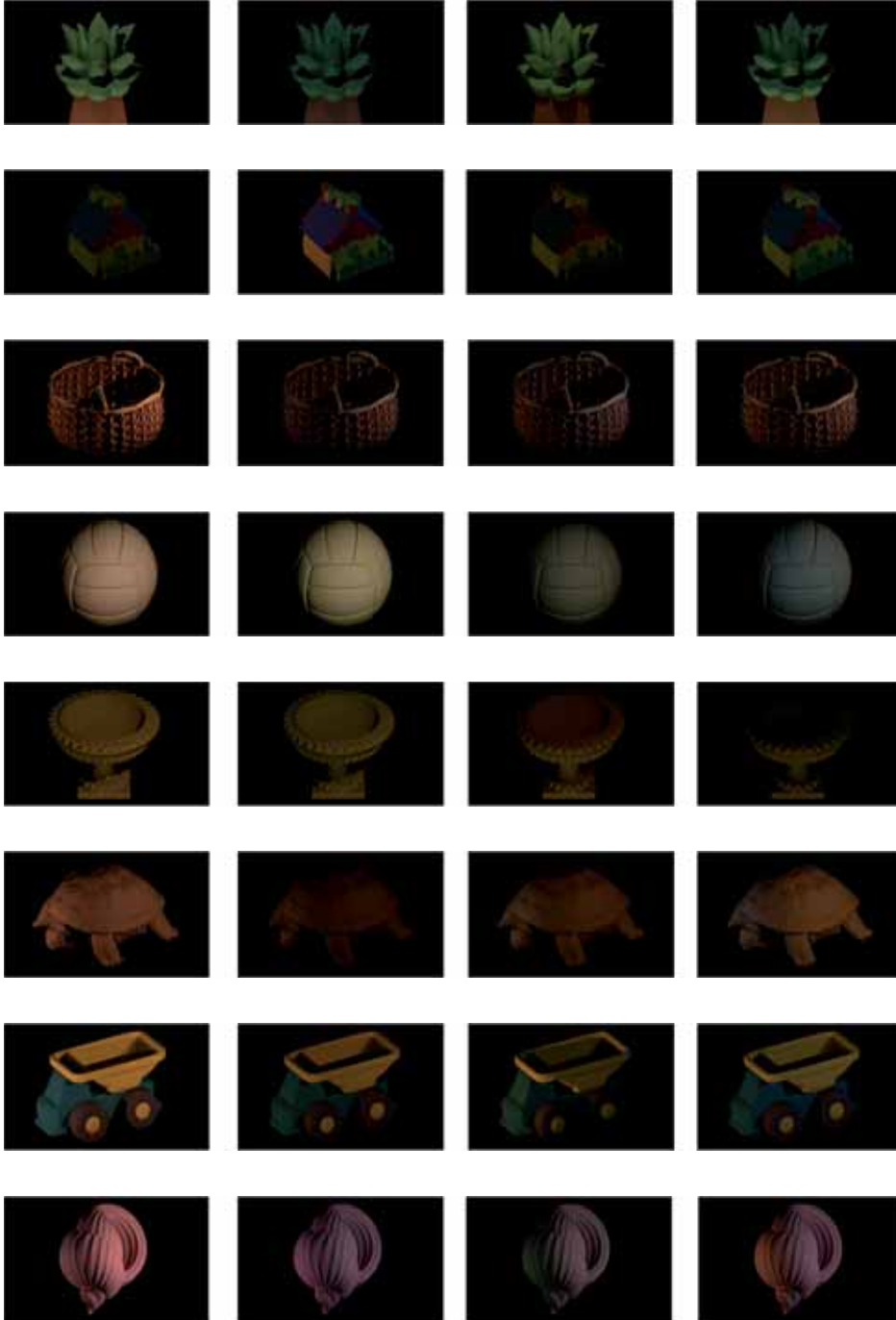


Figure C.1: Images of single objects in our synthetic dataset.

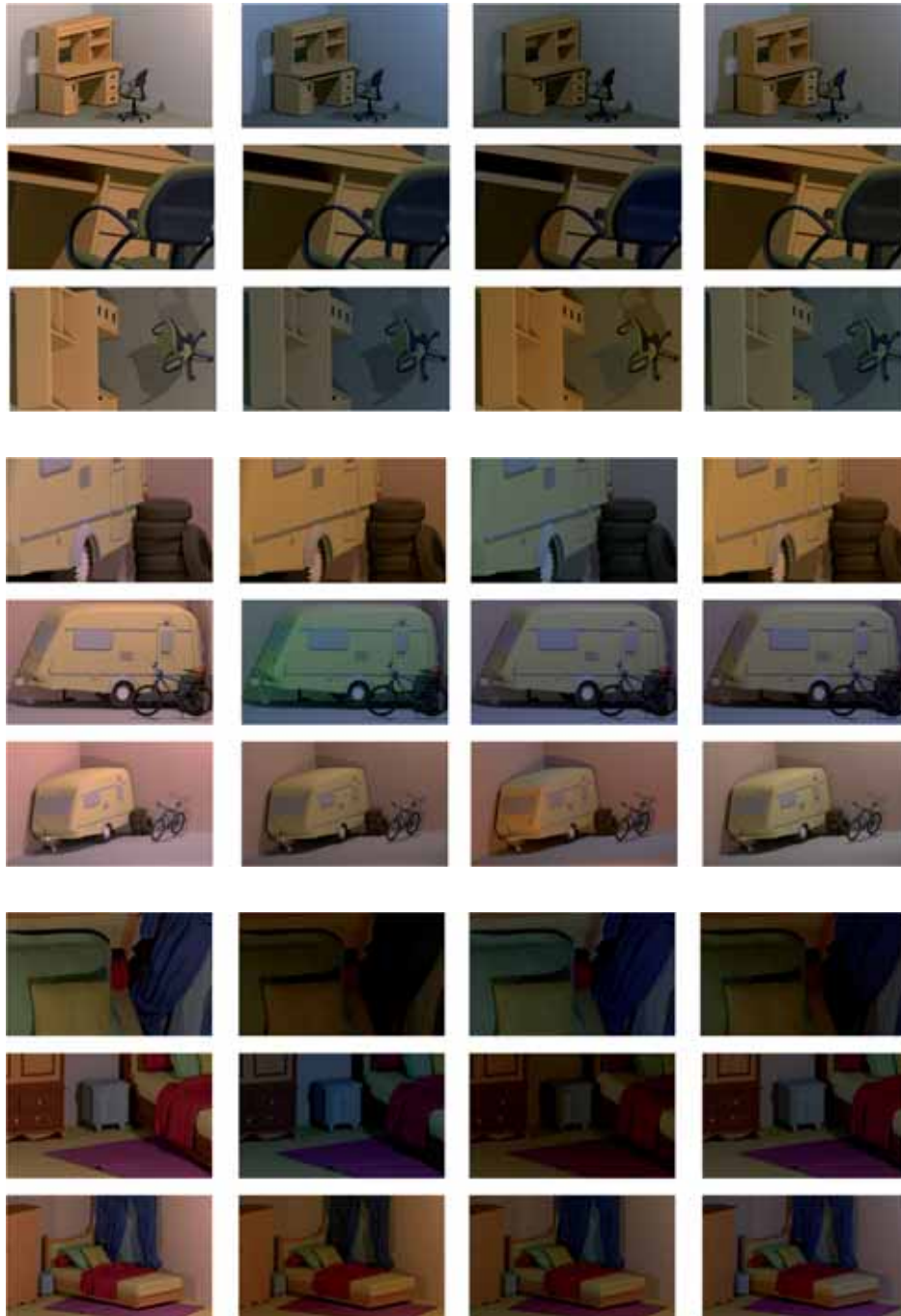


Figure C.2: Images of complex scenes in our synthetic dataset.



# Bibliography

- [1] <http://www.blender.org>.
- [2] <http://www.dxomark.com/>.
- [3] <http://www.flickr.com/>.
- [4] <http://www.uef.fi/spectral/spectral-database>.
- [5] <http://www.yafaray.org>.
- [6] <http://xritephoto.com/>.
- [7] IEC 61966-2-1 / FDIS: Multimedia systems and equipment - Colour measurement and management - Part 2-1: Colour management - Default RGB colour space - sRGB, 1999.
- [8] Vahid Abolghasemi and Alireza Ahmadyfard. Improved image enhancement method for license plate detection. In *International Conference on Digital Signal Processing*, pages 435–438, 2007.
- [9] Edward H. Adelson. Perceptual organization and the judgment of brightness. *Science*, 262(5142):2042–2044, 1993.
- [10] Edward H. Adelson. Lightness perception and lightness illusions. *The New Cognitive Neurosciences*, pages 339–351, 2000.
- [11] Edward H. Adelson and Alex P. Pentland. The perception of shading and reflectance. *Perception as Bayesian Inference*, pages 409–423, 1996.
- [12] Barton L. Anderson and Jonathan Winawer. Image segmentation and lightness perception. *Nature*, 434:79–83, 2005.
- [13] Alessandro Artusi, Francesco Banterle, and Dmitry Chetverikov. A survey of specular removal methods. *Computer Graphics Forum*, 30(8):2208–2230, 2011.
- [14] Shai Bagon. Matlab wrapper for graph cut, 2006.
- [15] Ricard Balagué. Exploring the combination of color cues for intrinsic image decomposition. Master’s thesis, Computer Vision Center - Universitat Autònoma de Barcelona, 2014.

- [16] Kobus Barnard and Brian Funt. Camera characterization for color research. *Color Research and Application*, 27(3):152–163, 2002.
- [17] Kobus Barnard, Lindsay Martin, Brian Funt, and Adam Coath. A data set for color research. *Color Research and Application*, 27(3):148–152, 2002.
- [18] Richard Barrett, Michael Berry, Tony F. Chan, James Demmel, June Donato, Jack Dongarra, Victor Eijkhout, Roldan Pozo, Charles Romine, and Henk van der Vorst. *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*. Society for Industrial and Applied Mathematics, 1994.
- [19] Jonathan T. Barron and Jitendra Malik. Color constancy, intrinsic images, and shape estimation. In *European Conference on Computer Vision*, pages 57–70, 2012.
- [20] Jonathan T. Barron and Jitendra Malik. Shape, albedo, and illumination from a single image of an unknown object. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 334–341, 2012.
- [21] Jonathan T. Barron and Jitendra Malik. Intrinsic scene properties from a single rgb-d image. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 17–24, 2013.
- [22] Jonathan T. Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.
- [23] Harry G. Barrow and J. Martin Tenenbaum. Recovering intrinsic scene characteristics from images. In *Computer Vision Systems*, pages 3–26, 1978.
- [24] Shida Beigpour, Marc Serra, Joost van de Weijer, Robert Benavente, Maria Vanrell, and Dimitris Samaras. Intrinsic image evaluation on synthetic complex scenes. In *International Conference on Image Processing*, pages 285–289, 2013.
- [25] Shida Beigpour and Joost van de Weijer. Object recoloring based on intrinsic image estimation. In *International Conference on Computer Vision*, pages 327–334, 2011.
- [26] Peter N. Belhumeur, David J. Kriegman, and Alan L. Yuille. The bas-relief ambiguity. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1060–1066, 1997.
- [27] Matt Bell and William T. Freeman. Learning local evidence for shading and reflection. In *International Conference on Computer Vision*, pages 670–677, 2001.
- [28] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics*, 33(4):159:1–159:12, 2014.
- [29] Robert Benavente, Maria Vanrell, and Ramon Baldrich. A data set for fuzzy colour naming. *Color Research and Applications*, 31(1):48–56, 2006.



- [30] Robert Benavente, Maria Vanrell, and Ramon Baldrich. Parametric fuzzy sets for automatic color naming. *Journal of the Optical Society of America A*, 25(10):2582–2593, 2008.
- [31] Brent Berlin and Paul Kay. *Basic Color Terms: Their Universality and Evolution*. University of California Press, Berkeley, 1969.
- [32] Adrien Bousseau, Sylvain Paris, and Frédo Durand. User assisted intrinsic images. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2009)*, 28(5):1–10, 2009.
- [33] Yuri Boykov, Olga Veksler, and Ramin Zabih. Efficient approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1222–1239, 2001.
- [34] William L. Briggs, Van Emden Henson, and Steve F. McCormick. *A Multigrid Tutorial (2nd ed.)*. Society for Industrial and Applied Mathematics, 2000.
- [35] Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision*, pages 611–625, 2012.
- [36] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *Journal of Scientific Computing*, 16(5):1190–1208, 1995.
- [37] Jason Chang, Randi Cabezas, and John W. Fisher III. Bayesian nonparametric intrinsic image decomposition. In *European Conference on Computer Vision*, pages 704–719, 2014.
- [38] Qifeng Chen and Vladlen Koltun. A simple model for intrinsic image decomposition with depth cues. In *International Conference on Computer Vision*, pages 241–248, 2013.
- [39] Leandro Cruz, Djalma Lucio, and Luiz Velho. Kinect and rgbd images: Challenges and applications. In *SIBGRAPI Conference on Graphics, Patterns and Images Tutorials*, pages 36–49, 2012.
- [40] Benjamin A. Darling, James A. Fewerda, Roy S. Berns, and Tongbo Chen. Real-time multispectral rendering with complex illumination. In *19th Color and Imaging Conference*, pages 345–351, 2010.
- [41] Maryam M. Darrodi, Graham Finlayson, Teresa Goodman, and Michal Mackiewicz. Reference data set for camera spectral sensitivity estimation. *Journal of the Optical Society of America A*, 32(3):381–391, 2015.
- [42] Yue Dong, Xin Tong, Fabio Pellacini, and Baining Guo. Appgen: Interactive material modeling from a single image. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2011)*, 30(6):1–10, 2011.

- [43] Richard Dosselmann and Xue D. Yang. A comprehensive assessment of the structural similarity index. *Signal, Image and Video Processing*, 5.
- [44] Mark S. Drew and Graham D. Finlayson. Multispectral processing without spectra. *Journal of the Optical Society of America A*, 20(7):1181–1193, 2003.
- [45] Jean-Denis Durou, Maurizio Falcone, and Manuela Sagona. Numerical methods for shape-from-shading: A new survey with benchmarks. *Computer Vision and Image Understanding*, 109(1):22–43, 2008.
- [46] Michael D’Zmura and Peter Lennie. Mechanisms of color constancy. *Journal of the Optical Society of America A*, 3(10):1662–1672, 1986.
- [47] Vebjørn Ekroll, Franz Faul, and Reinhard Niederée. The peculiar nature of simultaneous colour contrast in uniform surrounds. *Vision Research*, 44(15):1765–1786, 2004.
- [48] Graham Finlayson, Mark Drew, and Brian Funt. Color constancy: Generalized diagonal transforms suffice. *Journal of the Optical Society of America A*, 11(11):3011–3019, 1994.
- [49] Graham Finlayson, Mark Drew, and Cheng Lu. Intrinsic images by entropy minimization. In *European Conference on Computer Vision*, pages 582–595, 2004.
- [50] Graham D. Finlayson, Mark S. Drew, and Brian V. Funt. Spectral sharpening: Sensor transformations for improved color constancy. *Journal of the Optical Society of America A*, 11(5):1553–1563, 1994.
- [51] Graham D. Finlayson, Steven D. Hordley, and Paul M. Hubel. Color by correlation: A simple, unifying framework for color constancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1209–1221, 2001.
- [52] Graham D. Finlayson and Elisabetta Trezzi. Shades of gray and colour constancy. In *SID Color Imaging Conference*, pages 37–41, 2004.
- [53] David Forsyth. A novel algorithm for color constancy. *International Journal of Computer Vision*, 5(1):5–35, 1990.
- [54] David H. Foster. Color constancy. *Vision Research*, 51(7):674–700, 2011.
- [55] William T. Freeman and Paul A. Viola.
- [56] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.
- [57] Brian Funt, Mark Drew, and Michael Brockington. Recovering shading from color images. In *European Conference on Computer Vision*, pages 124–132, 1992.

- [58] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, 2010.
- [59] Elena Garces, Adolfo Munoz, Jorge Lopez-Moreno, and Diego Gutierrez. Intrinsic images by clustering. *Computer Graphics Forum (Proceedings of the Eurographics Symposium on Rendering)*, 31(4):1415–1424, 2012.
- [60] Peter V. Gehler, Carsten Rother, Martin Kiefel, Lumin Zhang, and Bernhard Schölkopf. Recovering intrinsic images with a global sparsity prior on reflectance. In *Neural Information Processing Systems*, pages 765–773, 2011.
- [61] Athinodoros S. Georghiadis, Peter N. Belhumeur, and David Kriegman.
- [62] Theo Gevers, Arjan Gijsenij, Joost van de Weijer, and Jan-Mark Geusebroek. *Color in Computer Vision : Fundamentals and Applications*. The Wiley-IS&T, 2012.
- [63] Arjan Gijsenij, Theo Gevers, and Joost van de Weijer. Computational color constancy: Survey and experiments. *IEEE Transactions on Image Processing*, 20(9):2475–2489, 2011.
- [64] Alan Gilchrist. The perception of surface blacks and whites. *Scientific American*, 24(3):88–97, 1979.
- [65] Alan Gilchrist, Stanley Delman, and Alan Jacobsen. The classification and integration of edges as critical to the perception of reflectance and illumination. *Perception & Psychophysics*, 33(5):425–436, 1983.
- [66] Alan Gilchrist, Stanley Delman, and Alan Jacobsen. Lightness contrast and failures of constancy: A common explanation. *Perception & Psychophysics*, 43(5):415–424, 1988.
- [67] Ernst H. Gombrich. *The Story of Art*. Phaidon Press, 1995.
- [68] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing (3rd Edition)*. Prentice Hall, 2008.
- [69] Meng M. Graham D. J. Artistic representations: clues to efficient coding in human vision. *Visual Neuroscience*, 28(4):371–379, 2011.
- [70] Lewis D. Griffin. Optimality of the basic colour categories for classification. *Journal of the Royal Society Interface*, 3(6):71–85, 2006.
- [71] Roger Grosse, Micah K. Johnson, Edward H. Adelson, and William T. Freeman. Ground-truth dataset and baseline evaluations for intrinsic image algorithms. In *International Conference on Computer Vision*, pages 2335–2342, 2009.
- [72] Frederick Hartt and David Wilkins. *History of Italian Renaissance Art (7th Edition)*. Pearson, 2010.

- [73] W. Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.
- [74] Michael T. Heath, Esmond Ng, and Barry W. Peyton. Parallel algorithms for sparse linear systems. *SIAM Review*, 33(3):420–460, 1991.
- [75] Hermann von Helmholtz. *Treatise on Physiological Optics, Vol. II*. 1866.
- [76] Ewald Hering. *Outlines of a Theory of Light Sense*. 1874.
- [77] Magnus R. Hestenes and Eduard Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49(6):409–436, 1952.
- [78] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.
- [79] Steven D. Hordley and Graham D. Finlayson. Reevaluation of color constancy algorithm performance. *Journal of the Optical Society of America A*, 23(5):1008–1020, 2006.
- [80] Berthold K.P. Horn. Determining lightness from an image. *Computer Graphics and Image Processing*, 3(4):277–299, 1974.
- [81] Berthold K.P. Horn. Obtaining shape from shading information. *The Psychology of Machine Vision*, pages 115–155, 1975.
- [82] Berthold K.P. Horn. Understanding image intensities. *Artificial Intelligence*, 8(2):201–231, 1977.
- [83] Pang-Ting Huang, Yi-Ming Chan, Li-Chen Fu, Shih-Shinh Huang, Pei-Yung Hsiao, Wei-Yu Wu, Chun-Cheng Lin, Kuo-Ching Chang, and Ping-Min Hsu. Pedestrian detection system in low illumination conditions through fusion of image and range data. In *International Conference on Intelligent Transportation Systems*, pages 2253–2254, 2014.
- [84] David H. Hubel and Torsten N. Wiesel. *Brain and Visual Perception: The Story of a 25-Year Collaboration*. Oxford University Press, 2004.
- [85] Paul M. Hubel, Jack Holm, Graham D. Finlayson, and Mark S. Drew. Matrix calculations for digital photography. In *Color Imaging Conference: Color Science, Systems, and Applications*, pages 105–111, 1997.
- [86] David S. Immel, Michael F. Cohen, and Donald P. Greenberg. A radiosity method for non-diffuse environments. *ACM SIGGRAPH*, 20(4):133–142, 1986.
- [87] Harold Jeffreys and Bertha S. Jeffreys. *Methods of Mathematical Physics (3rd ed.)*. Cambridge University Press, 1988.
- [88] Henrik W. Jensen. *Realistic image synthesis using photon mapping*. A.K. Peters, Ltd., Natick, MA, USA, 2001.

- [89] Junho Jeon, Sunghyun Cho, Xin Tong, and Seungyong Lee. Intrinsic image decomposition using structure-texture separation and surface normals. In *European Conference on Computer Vision*, pages 218–233, 2014.
- [90] Xiaoyue Jiang, Andrew J. Schofield, and Jeremy L. Wyatt. Correlation-based intrinsic image extraction from a single image. In *European Conference on Computer Vision*, pages 58–71, 2010.
- [91] Deane B. Judd. Color in business science and industry. *Applied Spectroscopy*, 7(2):90–91, 1953.
- [92] Peter K. Kaiser and Robert M. Boynton. *Human Color Vision (2nd Edition)*. Optical Society of America, 1996.
- [93] Rei Kawakami, Hongxun Zhao, Robby T. Tan, and Katsushi Ikeuchi. Camera spectral sensitivity and white balance estimation from sky images. *International Journal of Computer Vision*, 105(3):187–204, 2013.
- [94] Seung-Jean Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. An interior-point method for large-scale l1-regularized least squares. *Selected Topics in Signal Processing*, 1(4):606–617, 2007.
- [95] Frederick A.A. Kingdom. Perceiving light versus material. *Vision Research*, 48(20):2090–2105, 2008.
- [96] Frederick A.A. Kingdom. Lightness, brightness and transparency: A quarter century of new ideas, captivating demonstrations and unrelenting controversy. *Vision Research*, 51(7):652–673, 2011.
- [97] Gudrun J. Klinker. *A Physical Approach to Color Image Understanding*. A. K. Peters, Ltd., 1993.
- [98] Vladimir Kolmogorov. Convergent tree-reweighted message passing for energy minimization. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1568–1583, 2006.
- [99] Naejin Kong, Peter V. Gehler, and Michael J. Black. Intrinsic video. In *European Conference on Computer Vision*, pages 360–375, 2014.
- [100] Philipp Kraehenbuehl and Vladlen Koltun. Parameter learning and convergent inference for dense random fields. In *International Conference on Machine Learning*, pages 513–521, 2013.
- [101] James M. Kraft and David H. Brainard. Mechanisms of color constancy under nearly natural viewing. *Proceedings of the National Academy of Sciences of the United States of America*, 96(1):307–312, 1999.
- [102] Pierre-Yves Laffont, Adrien Bousseau, Sylvain Paris, Frédo Durand, and George Drettakis. Coherent intrinsic images from photo collections. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2012)*, 31(6):1–11, 2012.

- [103] Edwin H. Land. The retinex theory of color vision. *Scientific American*, 237(6):108–128, 1977.
- [104] Edwin H. Land and John McCann. Lightness and retinex theory. *Journal of the Optical Society of America*, 61(1):1–11, 1971.
- [105] Kyong Joon Lee, Qi Zhao, Xin Tong, Minmin Gong, Shahram Izadi, Sang Uk Lee, Ping Tan, and Stephen Lin. Estimation of intrinsic image sequences from image+depth video. In *European Conference on Computer Vision*, pages 327–340, 2012.
- [106] Thomas Leung and Jitendra Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision*, 43(1):29–44, 2001.
- [107] Joerg Liebelt and Cordelia Schmid. Multi-view object class detection with a 3d geometric model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1688–1695, 2010.
- [108] Xiaopei Liu, Liang Wan, Yingge Qu, Tien-Tsin Wong, Stephen Lin, Chi-Sing Leung, and Pheng-Ann Heng. Intrinsic colorization. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2008)*, 27(5):1–9, 2008.
- [109] Alexander D. Logvinenko. Lightness induction revisited. *Perception*, 28(7):803–816, 1999.
- [110] Alexander D. Logvinenko, Edward H. Adelson, Deborah A. Ross, and David Somers. Straightness as a cue for luminance edge interpretation. *Perception & Psychophysics*, 67(1):120–128, 2005.
- [111] Barbara London, John Upton, and Jim Stone. *Photography (11th Edition)*. Pearson, 2014.
- [112] Zhi-Quan Luo and Paul Tseng. On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72(1):7–35, 1992.
- [113] Sean P. MacEvoy and Michael A. Paradiso. Lightness constancy in primary visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 98(15):8827–8831, 2001.
- [114] Will Maddern, Alex Stewart, Colin McManus, Ben Upcroft, Winston Churchill, and Paul Newman. Illumination invariant imaging: Applications in robust vision-based localisation, mapping and classification for autonomous vehicles. In *International Conference on Robotics and Automation*, 2014.
- [115] Daniel Malacara. *Color Vision and Colorimetry: Theory and Applications (2nd Edition)*. SPIE Press, 2011.
- [116] Javier Marin, David Vázquez, David Gerónimo, and Antonio M. López. Learning appearance in virtual scenarios for pedestrian detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 137–144, 2010.

- [117] David Marr. Representing visual information. *AI Memo-415, Artificial Intelligence Laboratory, MIT*, 1977.
- [118] Yasuyuki Matsushita, Stephen Lin, Sing Bing Kang, and Heung Shum. Estimating intrinsic images from image sequences with biased illumination. In *European Conference on Computer Vision*, pages 274–286, 2004.
- [119] Yasuyuki Matsushita, Ko Nishino, Katsushi Ikeuchi, and Masao Sakauchi. Illumination normalization with time-dependent intrinsic images for video surveillance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(10):1336–1347, 2004.
- [120] Fabio Metelli. The perception of transparency. *Scientific American*, 230(4):90–98, 1974.
- [121] Addy Ngan, Frédo Durand, and Wojciech Matusik. Experimental analysis of brdf models. In *Eurographics Conference on Rendering Techniques*, pages 117–126, 2005.
- [122] Adriana Olmos and Frederick A. Kingdom. A biologically inspired algorithm for the recovery of shading and reflectance images. *Perception*, 33(12):1463–1473, 2004.
- [123] Ido Omer and Michael Werman. Color lines: Image specific color representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 946–953, 2004.
- [124] Alejandro Parraga, Ramon Baldrich, and Maria Vanrell. Accurate mapping of natural scenes radiance to cone activation space: A new image dataset. In *European Conference on Colour in Graphics, Imaging, and Vision and International Symposium on Multispectral Colour Science*, pages 50–57, 2010.
- [125] Freeman Patterson and André Gallant. *Photo Impressionism and the Subjective Image*. Key Porter Books, 2001.
- [126] Matt Pharr and Greg Humphreys. *Physically Based Rendering: From Theory to Implementation*. The Morgan Kaufmann series in interactive 3D technology. Elsevier Science, 2010.
- [127] Alistair W. G. Pike, Dirk L. Hoffmann, Marcos Garcia-Diez, Paul B. Pettitt, José J. Alcolea, Rodrigo De Balbín, César Gonzalez-Sainz, Carmen de las Heras, Jose A. Lasheras, Ramón Montes, and João Zilhão. U-series dating of paleolithic art in 11 caves in Spain. *Science*, 336(6087):1409–1413, 2012.
- [128] Jose A. Rodriguez-Serrano, Harsimrat Sandhawalia, Raja Bala, Florent Perronnin, and Craig Saunders. Data-driven vehicle identification by image matching. In *Computer Vision—ECCV 2012. Workshops and Demonstrations*, pages 536–545, 2012.

- [129] Andrew F. Rossi and Michael A. Paradiso. Neural correlates of perceived brightness in the retina, lateral geniculate nucleus, and striate cortex. *Journal of Neuroscience*, 19(14):6145–6156, 1999.
- [130] Bilge Sayim and Patrick Cavanagh. The art of transparency. *i-Perception*, 2(7):679–696, 2015.
- [131] János Schanda. *Colorimetry: Understanding the CIE System*. Wiley Interscience, 2007.
- [132] Heinrich Schfer. *Principles of Egyptian Art*. Griffith Institute, 1986.
- [133] Marc Serra, Olivier Penacchio, Robert Benavente, and Maria Vanrell. Names and shades of color for intrinsic image estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 278–285, 2012.
- [134] Marc Serra, Olivier Penacchio, Robert Benavente, Maria Vanrell, and Dimitris Samaras. The photometry of intrinsic images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1494–1501, 2014.
- [135] Steven A. Shafer. Using color to separate reflection components. *Color Research and Application*, 10(4):210–218, 1985.
- [136] Jianbing Shen, Xiaoshan Yang, Yunde Jia, and Xuelong Li. Intrinsic images using optimization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3481–3487, 2011.
- [137] Li Shen, Ping Tan, and Stephen Lin. Intrinsic image decomposition with non-local texture cues. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2008.
- [138] Li Shen and Chuohao Yeo. Intrinsic images decomposition using a local and global sparse representation of reflectance. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 697–704, 2011.
- [139] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgb-d images. In *European Conference on Computer Vision*, pages 746–760, 2012.
- [140] Eero P. Simoncelli and William T. Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *IEEE International Conference on Image Processing*, pages 444–447, 1995.
- [141] Pawan Sinha and Edward Adelson. Recovering reflectance and illumination in a world of painted polyhedra. In *International Conference on Computer Vision*, pages 156–163, 1993.
- [142] Deqing Sun, Stefan Roth, and Michael J. Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *International Journal of Computer Vision*, 106(2):115–137, 2014.



- [143] Robby T. Tan and Katsushi Ikeuchi. Intrinsic properties of an image with highlights. In *Meeting on Image Image Recognition and Understanding*, pages 465–470, 2004.
- [144] Yichuan Tang, Ruslan Salakhutdinov, and Geoffrey Hinton. Deep lambertian networks. In *International Conference on Machine Learning*, pages 1623–1630, 2012.
- [145] Marshall F. Tappen, Edward H. Adelson, and William T. Freeman. Estimating intrinsic component images using non-linear regression. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1992–1999, 2006.
- [146] Marshall F. Tappen, William T. Freeman, and Edward H. Adelson. Recovering intrinsic images from a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(9):1459–1472, 2005.
- [147] Martin J. Tovée. *An Introduction to the Visual System*. Cambridge University Press, 1996.
- [148] Quoc Tran-Dinh, Ion Necoara, and Moritz Diehl. Fast inexact distributed optimization algorithms for separable convex optimization. *Journal of Optimization*.
- [149] Joost van de Weijer and Arjan Gijsenij. Edge-based color constancy. *IEEE Transactions on Image Processing*, 16(9):2207–2214, 2007.
- [150] David Vázquez, Antonio M. López, and Daniel Ponsa. Unsupervised domain adaptation of virtual and real worlds for pedestrian detection. In *International Conference on Pattern Recognition*, pages 3492–3495, 2012.
- [151] Eduard Vazquez, Ramon Baldrich, Joost Van de Weijer, and Maria Vanrell. Describing reflectances for colour segmentation robust to shadows, highlights and textures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):917–930, 2011.
- [152] Javier Vazquez-Corral and Marcelo Bertalmio. Spectral sharpening of color sensors: Diagonal color constancy and beyond. *Sensors*, 14(3):3965–3985, 2014.
- [153] Javier Vazquez-Corral, David Connah, and Marcelo Bertalmio. Perceptual color characterization of cameras. *Sensors*, 14(12):23205–23229, 2014.
- [154] Vibhav Vineet, Carsten Rother, and Philip H.S. Torr. Higher order priors for joint intrinsic image, objects, and attributes estimation. In *Neural Information Processing Systems*, pages 1–9, 2013.
- [155] Brian A. Wandell. *Foundations of vision*. 1995.
- [156] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- [157] Yair Weiss. Deriving intrinsic images from image sequences. In *International Conference on Computer Vision*, pages 68–75, 2001.

- [158] Gerhard West and Michael H. Brill. Necessary and sufficient conditions for von kries chromatic adaptation to give color constancy. *Journal of Mathematical Biology*, 15(2):249–258, 1982.
- [159] Genzhi Ye, Elena Garces, Yebin Liu, Qionghai Dai, and Diego Gutierrez. Intrinsic video and applications. *ACM Transactions on Graphics*, 33(4):80:1–80:11, 2014.
- [160] Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. Generalized belief propagation. In *Neural Information Processing Systems*, pages 689–695, 2000.
- [161] Qi Zhao, Ping Tan, Qiang Dai, Li Shen, Enhua Wu, and Stephen Lin. A closed-form solution to retinex with nonlocal texture constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1437–1444, 2012.

# List of Publications

## International Conferences

- Marc Serra, Olivier Penacchio, Robert Benavente, Maria Vanrell, Dimitris Samaras “The Photometry of Intrinsic Images”, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1494-1501, 2014.
- Shida Beigpour, Marc Serra, Joost van de Weijer, Robert Benavente, Maria Vanrell, Olivier Penacchio, Dimitris Samaras “Intrinsic Image Evaluation On Synthetic Complex Scenes”, *IEEE International Conference on Image Processing*, pp. 285-289, 2013.
- Marc Serra, Olivier Penacchio, Robert Benavente, Maria Vanrell, “Names and Shades of Color for Intrinsic Image Estimation”, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 278-285, 2012.

## Master Thesis

- Marc Serra, “Estimating Intrinsic Images from Physical and Categorical Color Cues”, *Universitat Autònoma de Barcelona*, 2010.