N° d'ordre : D.U. ?
EDSPIC : ?

Université Clermont Auvergne

École Doctorale
Sciences Pour l'Ingénieur de Clermont-Ferrand

Thèse présentée par :
Souheil Hadj Said

Formation Doctorale :
Électronique et Système

en vue de l'obtention du grade de

Docteur d'Université

Spécialité : Vision pour la Robotique

Illumination Estimation from Specular Highlights in Mixed Reality with Application in Diminished Reality

Prof. Isabelle Bloch	Examinatrice
Dr. Thomas Corpetti	Rapporteur
Prof. Cédric Demonceaux	Rapporteur
Prof. Adrien Bartoli	Directeur de thèse
Dr. Mohamed Tamaazousti	Co-encadrant

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Dr. Mohamed TAMAAZOUSTI for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis.

I would also like to thank my thesis director Prof. Adrien BARTOLI for his insightful comments and suggestions, but also for his unconditional availability and constant encouragements.

I thank my fellow labmates for the stimulating discussions.

I would like to thank my family: my parents and my brothers for supporting me spiritually throughout my thesis. Last but not least, I would like to thank my wife Abir for her love and constant support. But most of all, thank you for being my best friend.

Abstract

Diminished Reality (DR) is a video editing technique that alters reality by removing certain objects. It can be used as a preliminary step in Augmented Reality to replace real objects by virtual ones with different sizes and shapes. It can also be used solely, for example, in the case of virtually emptying a furnished apartment. The general approach of DR consists in three main steps. First, an inpainting technique is applied to a target region in the image to coherently remove an object. The image corresponds to a keyframe of the video stream. Second, the resulting inpainted region is transmitted to the next frames of the video stream by copying pixel intensities with respect to the camera pose and scene geometry. This consists in estimating the camera orientation and position in 3D which can be obtained by a Simultaneous Localization and Mapping (SLAM) technique. Third, the target region is updated with respect to the lighting change in the scene.

In this thesis, we focused on the third step of the DR pipeline. Although many DR applications have been proposed in the literature, few are the ones who dealt with light change in the scene. Most of past work assumes that the surface is Lambertian and therefore perfectly diffuse. However, this is often not true, especially in indoor environments. By identifying specular highlights as the main cause for lighting change in the target region, we proposed two main approaches to address this problem.

First, we proposed a specularity propagation method applied to real-time DR. Using the DR pipeline mentioned earlier, we integrated an interpolation function based on Thin-Plate Splines (TPS) in order to estimate the change ratios of the pixel intensities in the target region. This function is constrained by a number of specularity properties to achieve a plausible reconstruction of the specular highlights in the video stream. Our approach was tested on several real-time videos and achieved coherent reproduction of specularities in the context of DR.

Second, we addressed the lighting problem in DR and AR as an inverse rendering problem. To do so, we analyzed the image components as described in light reflection models. In Computer Graphics, local illumination models such as Phong's are used to render synthetic images in real-time. In this case, the parameters of the model are set by the user as inputs along with the scene's geometry, the light source configuration and the camera pose.

However, in a Mixed Reality (MR) application, the parameters of the model are unknown and have to be set in concordance with the real image from the camera. So, in this case we want to solve an inverse local illumination problem where the input is the real image. The output is the model's parameters along with the light source configuration, the scene's geometry and the camera pose. In this thesis, we proposed an exhaustive evaluation of the well-posedness of this problem with a focus on the specular highlights. The camera pose and the scene's geometry are estimated using the SLAM approach and the rest of the unknown parameters are estimated by minimizing a photometric cost. We showed that we can invert a local illumination model from the observation of a single specular highlight. Therefore, in the context of AR and DR applications, we do not need to know the number of light sources in the scene a priori since each specularity is processed separately. This also opens many perspectives for similar inversion problems like camera localization.

Table of contents

List of figures

List of tables

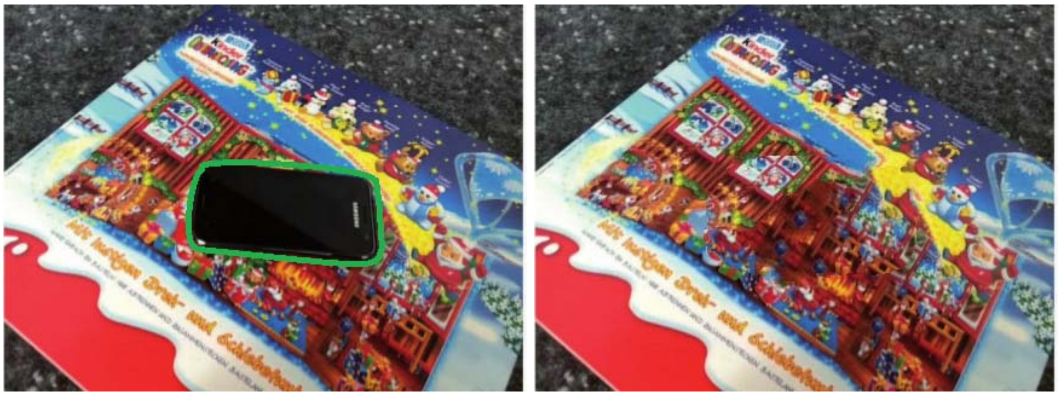
Chapter 1

Introduction

This thesis had been carried out between April 2015 and March 2018 in the LVIC (Laboratoire de Vision et Ingénierie des contenus) lab of CEA LIST, in Saclay, France and in Institut Pascal, UMR6602 CNRS, UCA, SIGMA in Clermont-Ferrand, France. This thesis has led to four international publications: IEEE Transactions on Visualization and Computer Graphics [40] (TVCG 2017), ACM Symposium on Virtual Reality Software and Technology [73] (VRST 2017), IET Image Processing [15] (2018), IEEE Eurographics [42] (2019) and two national publications: ORASIS, journées francophones des jeunes chercheurs en vision par ordinateur [41] (ORASIS 2017), Revue Française de Photogrammétrie et de Télédétection [14] (RFPT 2017).

1.1 Context

DR is a technique that alters a video stream in real-time to plausibly remove an undesired object [49, 28, 47, 56, 58, 68, 69, 78, 101]. This technique may be used in many applications. For example, some pieces of furniture may be removed to simulate different arrangements in a room [96, 109]. In live streams, advertising signs can be removed or replaced by new ones depending on the video. In Augmented Reality (AR) applications, markers are often used and they can be hidden to achieve seamless fusion between virtual objects and the real world [61, 65]. In many AR applications, a real object is replaced by a virtual one. DR is often used in these cases as a first step to achieve seamless blending of the virtual objects. Unlike video inpainting [7, 86], DR is a real-time application where the frames of a video stream are processed on the fly. So, we cannot use video frames as priors to reconstruct the real scene. The user would not know what is behind the undesired object anyway. So, our aim here is to generate a synthesized portion of the image that will plausibly blend with the real scene. A good result would require a coherent inpainting method that will spatially



(a) Original image (before inpainting) (b) Image after inpainting

Fig. 1.1 Illustration of the inpainting result (b) used in our DR pipeline on a keyframe image. The target region is delimited by a green line in the original image (a).

propagate the neighboring structure and texture inside the target region but also a coherent temporal propagation of this result into the next frames. The temporal propagation includes the integration of light change in the scene. If we consider the neighboring area surrounding the target region, the pixel intensities of this area are usually non constant when the camera moves due to the specular property of the surface. They are only constant when the surface is perfectly diffuse, which is a special case in the context of DR. In general, updating the target region depending on the lighting change is a crucial step in DR to ensure a good propagation of specular reflections and to obtain a plausible result. So, DR can be also seen as a spatio-temporal inpainting problem. A detailed description of its pipeline is presented in the next section. The second chapter represents our contribution to specularity propagation in the context of DR applications. In this thesis, we also focused on the lighting estimation for DR as well as AR applications. While propagating the specularity in the scene by comparing two frames can ensure a plausible approximation of the specular highlights, it will always require a visible portion of the specularity. The robust estimation of the scene's parameters with respect to a local illumination model can produce more accurate results regardless of the position of the specularity. In fact, it is possible to synthesize the missing region entirely using local illumination models. Generally speaking, the models consist of two components which are the diffuse component and the specular component. A third component is added in some models to approximate indirect reflections. it is called the ambient component. In the third chapter, we evaluate the well-posedness of this approach when constraining the input image to a single specular highlight region. The output is the synthetic image. It is obtained from the equation of the illumination model taking into account all the mentioned input parameters. The results of this study set the criteria for a robust estimation of the parameters of a local illumination model with a minimum amount of data. As a result, this approach can be applied to real-time AR and DR since we can estimate the model from a single keyframe.

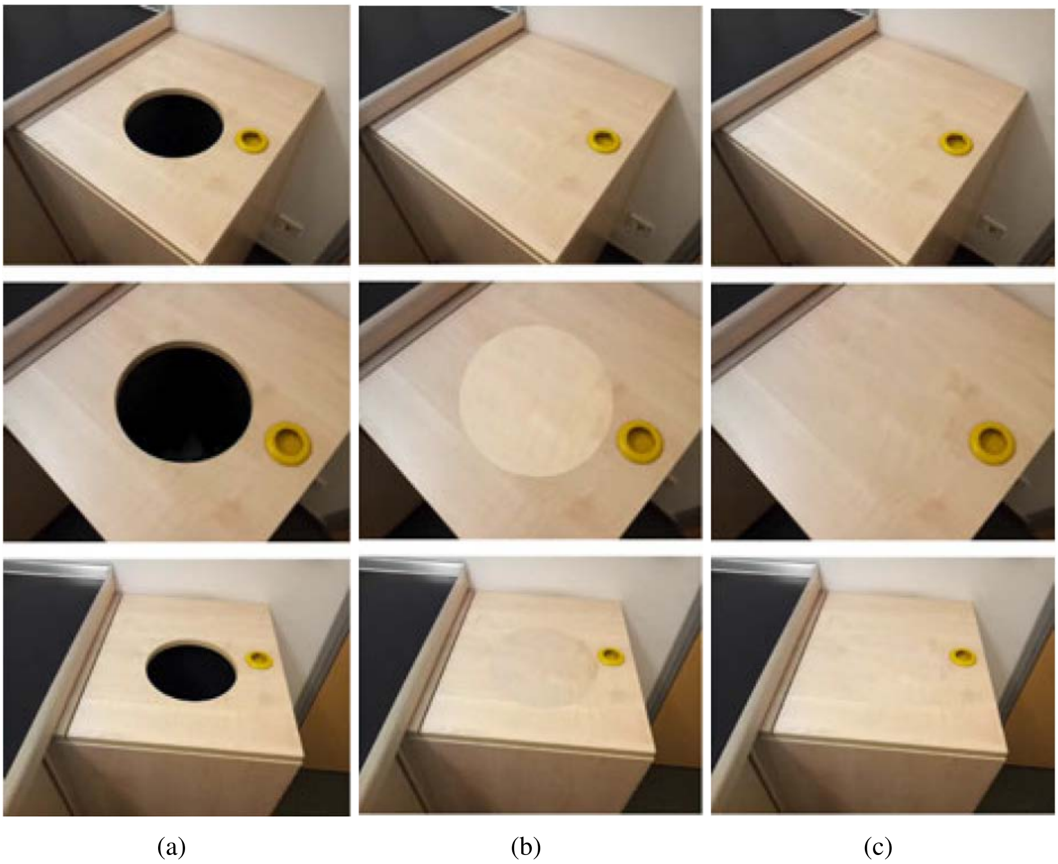


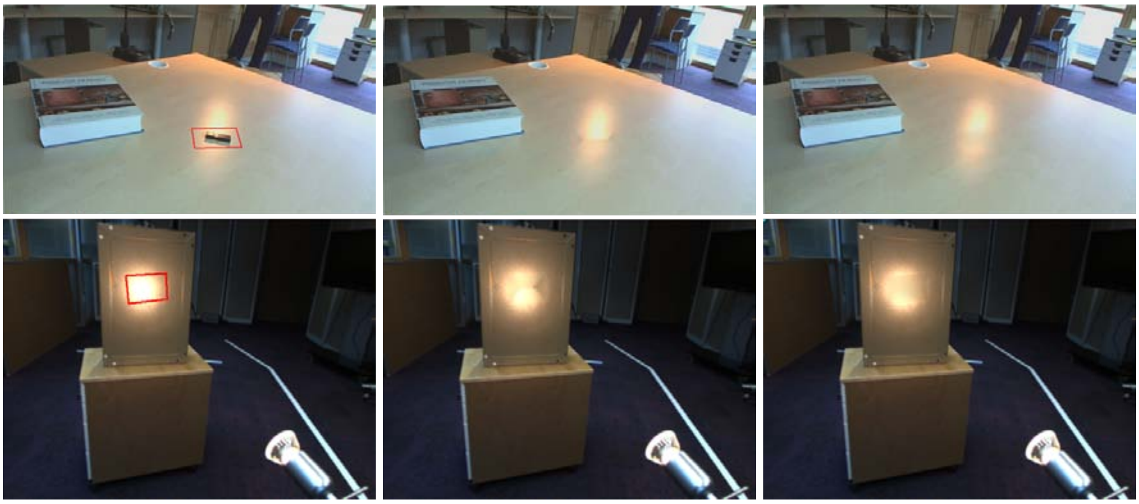
Fig. 1.2 Illustration of the light change correction in previous work of Herling *et al.* (a) is the original image (before inpainting). (b) is the DR result without light change update. (c) is the result of DR with light change update [48]

.

1.2 Pipeline

1.2.1 Image Inpainting

Image inpainting is the technique of coherently filling a missing region of the image with respect to the rest of the image. This was widely investigated in the literature for applications in image editing such as removing undesired objects from photos, adding special effects, blurring or older photo restoration. In DR applications, the missing region usually corresponds to a large portion of the image and the algorithm has to respect a very low time complexity. The state-of-the-art methods adapt the PatchMatch inpainting algorithm [3] which fits perfectly to the context of DR. PatchMatch proposes a fast matching algorithm between patches of the image. A randomized search is introduced to significantly reduce the computation time. While this may generate different results for the same input pair (image/target region), it has demonstrated plausible inpainting results. During this thesis, we used this as the inpainting method for our DR application. An example illustrating the efficiency of this method is reported in figure 1.1.



(a) Original image (b) DR by Herling *et al*[48] (c) DR by Kawai *et al* [58]

Fig. 1.3 Illustration of the limits of previous work on light change update for DR.

1.2.2 Camera Tracking

In DR applications, the camera usually moves in the scene. Therefore, we need to copy the inpainted region from the keyframe to the current frame with respect to the camera pose. The solution usually used for this problem is a Simultaneous Localization and Mapping algorithm (SLAM). It aims at determining simultaneously the camera pose and the scene geometry (which is referred to as mapping). In this thesis, we used the model-constrained SLAM approach proposed by Tamaazousti *et al* [99, 98]. In this approach, both known and unknown parts of the environment provide constraints on the camera motion. The known part is the 3D model of a known object inserted intentionally in the scene and the unknown part is the features in the rest of the scene. A nonlinear refinement process of an initial Structure-from-Motion (SfM) reconstruction takes advantage of these two types of constraints to provide an optimized estimation of the camera pose. Knowing the intrinsic parameters of the camera, we can compute the transformation matrix between the keyframe and the current frame based on the estimated extrinsic parameters (position and rotation of the camera). Then, we can copy all pixels of the target region using this transformation matrix. In our case, the target region is planar. Updating the target region in the context of DR is a crucial step that has been mastered by the AR community. However, another important aspect to achieving good DR results has been less studied in literature which is the light change in the target region.

1.2.3 Light Change Updating

As the camera moves, the lighting may change because of the reflective properties of the materials involved in the scene or when the camera's exposition parameter is not set as constant (see figure 1.2). Here, we are only interested in the first case (the exposition parameter is set to constant). Most of previous work propose to compute the average offset of pixel intensities between the keyframe and the current frame. Only pixels inside the neighboring region of the target region are considered. The pixels intensities inside the target region are incremented by this average offset to compensate the light change. More advanced techniques interpolate light change in the target region with respect to the neighboring region using a customized function [47] or with an interpolation grid [58]. More details on these techniques are reported in the next chapter. Although there are some methods that address this problem, it is still not completely solved (see figure 1.3). In fact, they only work with largely soft specular highlights (see figure 1.2). In this thesis, we focused on this problem in order to present a more complete solution and enhance the user experience in the context of DR.

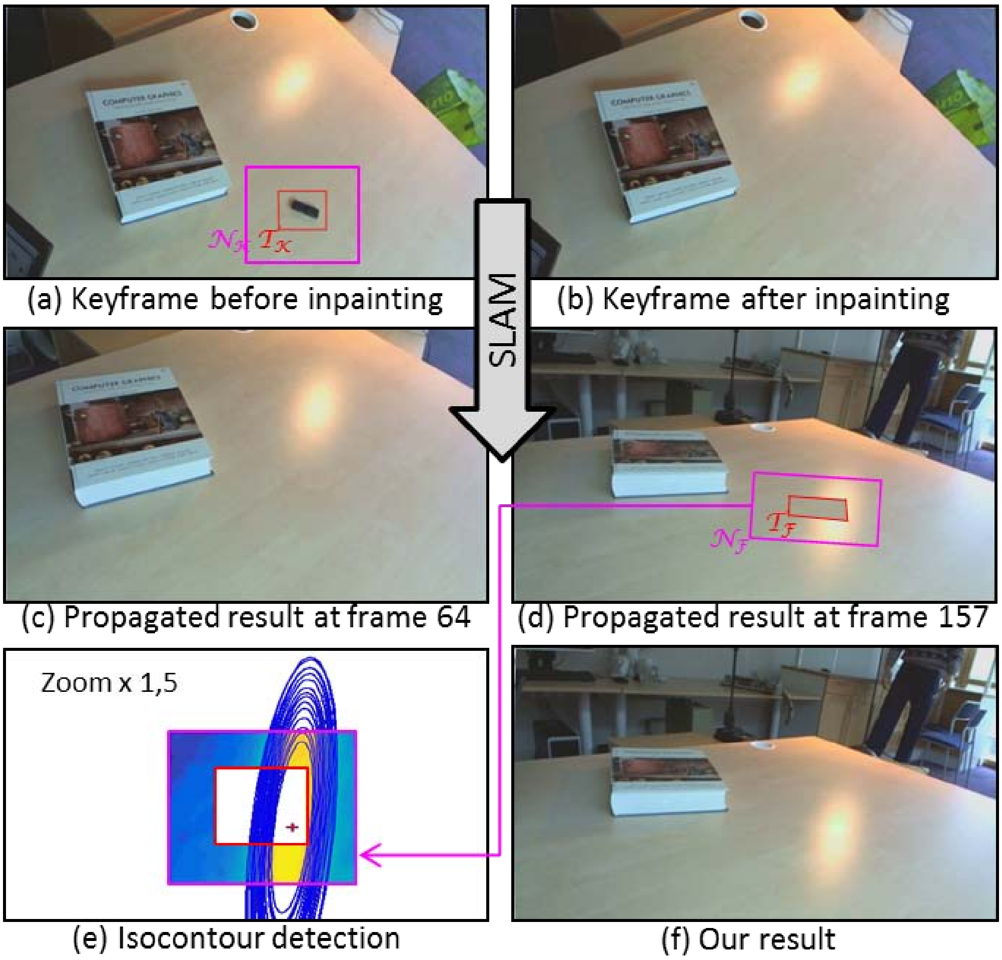
Chapter 2

Image-based Specularity Propagation for Diminished Reality

The work reported in this chapter was published in the IEEE Transactions on Visualization and Computer Graphics [40].

2.1 Introduction

In Diminished Reality applications, the user experience is largely enhanced by a realistic rendering quality. A good result has to ensure a seamless fusion between the real video and the synthesized part. The state-of-the-art image completion methods [3, 57, 45] allow a coherent replacement of the deleted region which blends perfectly with the rest of the image, even for textured surfaces. However, for temporal consistency in the video and for a real-time application, one cannot apply these methods in each frame. A solution is to use these methods on a selected image which we call the “keyframe". For the next frames of the video stream, one simply needs to copy the inpainting result, considering the camera movement as well as the illumination change around the target area. Here, we address the illumination change problem, which is a crucial stage in any DR pipeline. This is a difficult problem due to the complex nature of light reflections in the presence of glossy surfaces. In fact, illumination variation is often observed when the viewpoint changes. In these situations, any artifacts can significantly deteriorate the user's experience. Only two of the previous DR methods explicitly address this problem [49, 59]. We show that even in simple scenarios, with planar surfaces and a single point light source, these methods produce unconvincing results (see figure 2.4). This is because they only assume the continuity of the illumination variation. In this paper, we analyze the origin of this problem and show that it originates from the



(a) Keyframe before inpainting

(b) Keyframe after inpainting

(c) Propagated result at frame 64

(d) Propagated result at frame 157

(e) Isocontour detection

(f) Our result

$\mathcal{N}_\mathcal{K}$
$\mathcal{T}_\mathcal{K}$
SLAM
$\mathcal{N}_\mathcal{F}$
$\mathcal{T}_\mathcal{F}$
Zoom x 1,5

Fig. 2.1 Specularity propagation in DR. In (a), at frame 1 (the keyframe), the user selects a target region $\mathscr{T}_K$. In (b), we copy patches from the rest of the image to obtain the inpainted keyframe. In (c), the result is transformed using SLAM to frame 64 without specularity propagation. In (d), the transformed result at frame 157 is not visually convincing because a specularity is around $\mathscr{T}_K$. In (e), at frame 157, the image is rectified to the keyframe's image plane and the isocontours of light intensity are fitted with ellipses. This information is used by our model, whose output is shown in (f), to synthesize the specularity.

specular component in the image. Actually, local illumination models such as Phong's [89], Blinn-Phong's [10] and Cook-Torrance's [22] confirm that only the specular component of light reflection depends on the viewpoint. Moreover, it was shown that the specularities play a key role in scene perception by the human brain [9]. Therefore, we formulate this problem into a specularity propagation problem. More specifically, we consider the case of deleting an object lying on a planar specular surface illuminated by a point light source. From a set of real videos under these assumptions, we observe the structural properties of specularities and propose two new models:

- Our first model is called Smooth Propagation Model (SPM). It is generic and exploits the continuity and smoothness of light intensity. We use the *Thin-Plate Spline* (TPS) as a smooth function representation. Intensity's smoothness was exploited by Kim *et al.* [62] to separate the specular and diffuse components in a single image. SPM achieves state-of-the-art performances and works for general case scenarios, but incorporates few structural properties, and has similarities to previous work [49, 59].

- Our second model is called Constrained Propagation Model (CPM). It incorporates the observed structural properties of the specularity. It extends the first model by imposing additional structural constraints: the ellipticity of the intensity isocontours and the existence of a unique maximum intensity within a specularity. We refer to this as the *brightest point*. CPM is more specific to our assumptions but gives better results than previous methods.

Section 2 of this chapter describes the main structural properties of a specularity. Section 3 formally states the problem we aim to solve. Section 4 reviews previous solutions. Section 5 introduces our proposed models and algorithms. Finally, section 6 shows and discusses our experimental results.

2.2 The Structural Properties of a Specularity

By observing images of specularities on planar surfaces such as the ones in figure 2.2, we established some structural properties of a specularity. These are described in terms of how the light intensity behaves across a specularity:

1. **Smoothness.** The light variation is smooth, and thus continuous.

2. **Brightest point.** The specularity has a single brightest point located approximately at its center.

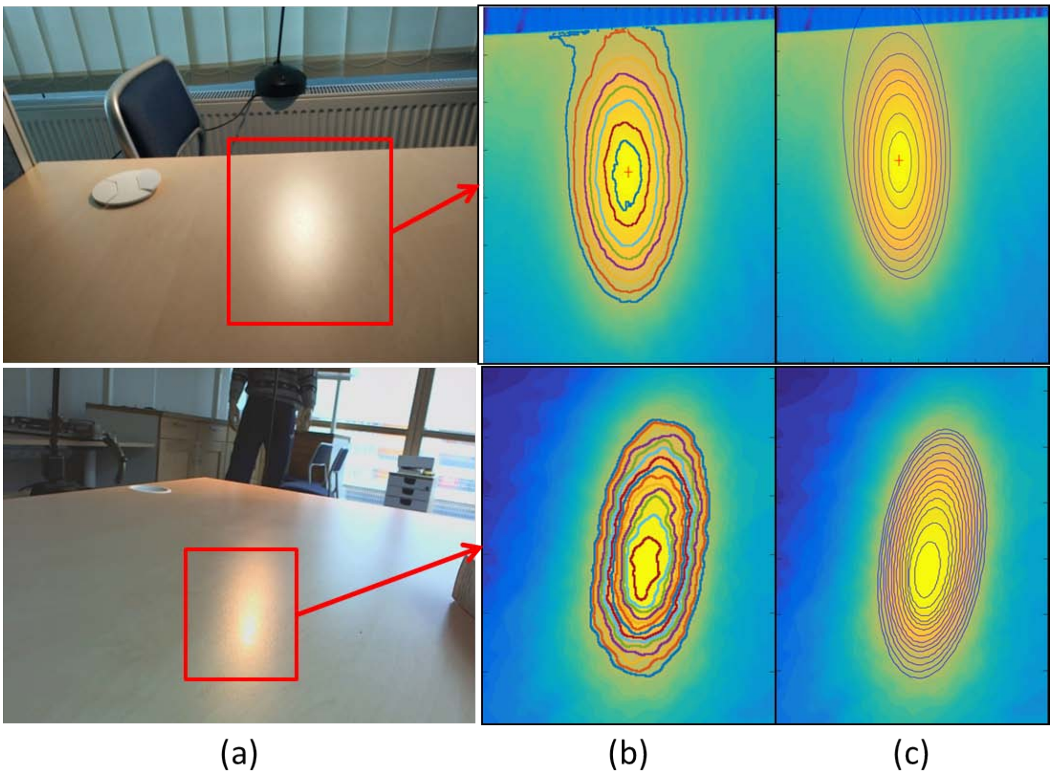


Fig. 2.2 In (a) the images show specular highlights on a flat surface. In (b) we show the corresponding light map around the specularity for each image and the intensity's isocontours. In (c), we show the fitted ellipses for these isocontours.

3. **Ellipticity.** The isocontours of a specularity are approximately elliptic.
4. **Monotonicity.** The further away the brightest point, the lower the intensity. This implies that the isocontours do not intersect.
5. **Additivity.** Following the local illumination models, the specular component is a term added to the ambient and diffuse terms.

Some of these properties were theoretically and empirically verified on models from Computer Graphics (specifically Phong's [89] and Blinn-Phong's [10]). In particular, it has been empirically verified in [83] that the elliptic shape is a good approximation for the specularity's isocontours in practice. The fifth property is directly deduced from the Phong illumination model, which suggests that the color intensity I at a given point is expressed as the sum of three components:

$$I = I_{\text{ambient}} + I_{\text{diffuse}} + I_{\text{specular}}. \tag{2.1}$$

These properties have not been considered for propagating specularities in DR in existing methods [47, 49, 59]. Our goal is to exploit them in order to improve the realism of specularity rendering in DR.

2.3 Background and Problem Statement

2.3.1 Notation

Scalars are in italics (*e.g.* x), vectors in bold upright (*e.g.* $\mathbf{v}$) and matrices in sans-serif (*e.g.* M). The elements of a vector are written as in $\mathbf{a}^\top = (a_1\ a_2\ a_3)$ where $^\top$ is vector and matrix transpose. The coordinates of a point in the image are written with a 2-vector $\mathbf{q}^\top = (x\ y)$. An image domain is written in uppercase calligraphic (*e.g.* $\mathscr{R}$). A group of points is written with uppercase italic (*e.g.* B) and the number of points in a group as $|B|$. Functions are written in upright Greek letters (e.g. ψ) or Latin lower case in italics. The Euclidean distance between two pixels $\mathbf{p}$ and $\mathbf{q}$ is denoted $d(\mathbf{p},\mathbf{q})$.

2.3.2 Problem Statement

Context

In this section, we introduce two major techniques used for DR. First, we explain the image inpainting technique. An image can be mathematically defined by a function χ giving the color intensities as:

$$\chi : \left\| \begin{array}{l} \mathscr{O} \subset \mathbb{R}^2 \to \mathbb{R}^n \\ \mathbf{p} \to \chi(\mathbf{p}), \end{array} \right. \tag{2.2}$$

where $\mathbf{p}$ represents a vector indicating the spatial coordinates of a pixel. For an *RGB* color space ($n = 3$), the image is described by three color intensity functions. So, χ can be written as $\chi^\top = (\chi^R\ \ \chi^G\ \ \chi^B)$. Image inpainting [38] was introduced as a term by Bertalmio *et al.* [8]. Since then, many real-time image inpainting techniques were proposed [7, 2, 20, 3, 23, 46, 57, 100]. In general, in the inpainting problem, the image described by χ (*i.e.* corresponding to each color channel of the image) is assumed to have gone through a degradation operation. As a result, the generic definition domain $\mathscr{O}$ of the input image χ can be seen as composed of two parts $\mathscr{O} = \mathscr{S} \cup \mathscr{T}$, $\mathscr{S}$ being the intact part of the image (the source region) and $\mathscr{T}$ the deleted part of the image which we search to recover (the target region). The goal of inpainting is to estimate the color intensities of the pixels $\mathbf{p}$ located in the target region $\mathscr{T}$. As a final result, this technique reconstructs the inpainted image described by $\hat{\chi}$. The objective in terms of quality is that the recovered region looks natural to the human eye, and is as physically plausible as possible. Typical inpainting artifacts are unconnected edges, blur, and inconsistent pieces of texture (also called texture garbage).

The second technique we use is Simultaneous Localization and Mapping (SLAM). In our work, it is used to localize the camera and therefore, map the target region in all the frames of the video stream. We denote a 3D point as $\mathbf{x} \in \mathbb{R}^3$, the rotation of the camera as $\mathsf{R} \in \mathbb{SO}(3)$ and its translation as $\mathbf{t} \in \mathbb{R}^3$. At each frame f, SLAM determines the coefficients of R and $\mathbf{t}$ that coherently projects a 3D point $\mathbf{x}$ to the camera's image plane. SLAM solves this problem in real time, and is available in mature software packages. We use the SLAM technique from [99].

DR as Spatio-Temporal Inpainting

We consider DR as a spatio-temporal inpainting problem. For the keyframe, spatial consistency is ensured by the inpainting technique. For the next frames of the video, SLAM propagates the spatially-consistent inpainting result while ensuring temporal consistency in the video. However, a specularity may appear around $\mathscr{T}$, which causes illumination variations. So, the spatial structure of the inpainting result should be properly modified in order to achieve spatio-temporal consistency. This modification is essential to obtain a realistic rendering result. We refer to the keyframe image by χ_K and the current video frame by χ_F. An inpainting technique is applied on χ_K to reconstruct the target region $\mathscr{T}_K$. We use a modified version of PatchMatch which is a real-time capable image inpainting approach initially proposed by Barnes *et al.* [3]. The inpainted image is then propagated to the next frames. We use SLAM to transform the current frame to the keyframe image plane. We therefore have dense pixel-wise correspondences between the target region in the keyframe and the one in the current frame. In other words, for each pixel $\mathbf{p}_K$ in the keyframe, we have a corresponding pixel $\mathbf{p}_F = \eta(\mathbf{p}_K)$, η being a homography function. So, we can transform the result of inpainting to all the frames of the video stream. We define the neighboring region $\mathscr{N}_K \subset \mathscr{S}$ centered around $\mathscr{T}_K$ with width $w_{\mathscr{N}} = z\, w_{\mathscr{T}}$ and height $h_{\mathscr{N}} = z\, h_{\mathscr{T}}$ (see figure 2.3). $\mathscr{N}$ is the set of neighboring pixels that are outside $\mathscr{T}$. $w_{\mathscr{T}}$ and $h_{\mathscr{T}}$ are, respectively, the width and the height of the target region selected by the user. $z > 1$ is set manually depending on the specularity's size to allow for an efficient observation of the specularity's isocontours. The larger the specularity, the greater z. In our experiments, we set $z = 2$. By observing the light variation in the current frame in $\mathscr{N}_F$, we aim to propagate this variation inside $\mathscr{T}_F$. For each pixel $\mathbf{p}_F$ in $\mathscr{N}_F$, the illumination variation coefficient is defined as:

$$v_{\mathbf{p}_K} = \chi_F(\mathbf{p}_F) - \chi_K(\mathbf{p}_K), \mathbf{p}_F \in \mathscr{N}_F. \tag{2.3}$$

For each frame f, knowing the variation coefficients of the pixels in $\mathscr{N}_F$, we aim to estimate the function ψ_F that returns the illumination variation for all pixels in $\mathscr{T}_F \cup \mathscr{N}_F$ and therefore, update their color intensities as:

$$\hat{\chi}_F(\mathbf{p}_F) = \psi_F(\mathbf{p}_K) + \hat{\chi}_K(\mathbf{p}_K),\ \mathbf{p}_F \in \mathscr{T}_F \cup \mathscr{N}_F. \tag{2.4}$$

$\psi_F(\mathbf{p}_K)$ can be seen as the estimated value of $v_{\mathbf{p}_F}$ if $\mathbf{p}_F \in \mathscr{T}_F$ and the real value of $v_{\mathbf{p}_F}$ if $\mathbf{p}_F \in \mathscr{N}_F$.

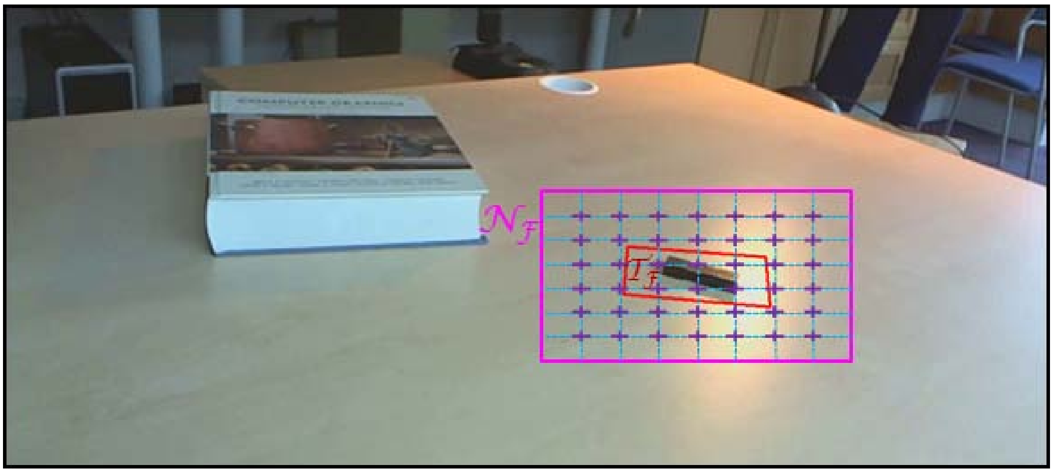


Fig. 2.3 Representing the region of interest in the current frame $\mathscr{F}$. The neighboring area $\mathscr{N}_\mathscr{F}$ is represented by the blue grid. The purple crosses represent the centers whose target values will be estimated from equation (2.12). The pixels inside the target region $\mathscr{T}_\mathscr{F}$ (delimited by a red contour) are then interpolated using the TPS.

2.4 State-of-the-Art

The literature has some real-time DR approaches. They use approximately the same pipeline as ours but with different image inpainting and camera tracking techniques [65, 68, 96]. However, only two of them consider the light change problem [49, 59]. They propose heuristic interpolation techniques to estimate the illumination variation in the target region. They use similar models, which suggest that the variation is continuous and smooth, and thus respect the first structural property of section 2. They however use different estimation approaches, explained in the next two sections.

2.4.1 Herling *et al.*

In the approach of Herling *et al.* [49], pixels from the boundary separating the target region from the rest of the image are monitored over time and the color difference at each of these pixels is computed. Then, a virtual grid $\mathscr{G}$ is defined, covering the target region. For each node of the grid inside the target region $\mathbf{c}$, the color correction function is determined by:

$$\psi_{\text{Herling}} : \left\| \begin{array}{l} \mathscr{G} \subset \mathscr{T}_F \rightarrow \mathbb{R} \\ \mathbf{c} \rightarrow \psi_{\text{Herling}}(\mathbf{c}), \end{array} \right. \tag{2.5}$$

where:

$$\psi_{\text{Herling}}(\mathbf{c}) = \frac{1}{\theta(\mathbf{c})} \sum_{j=1}^{|B_K|} (\chi_K(\mathbf{b}_{K,j}) - \chi_F(\mathbf{b}_{F,j}))e^{-\ |\mathbf{c}-\mathbf{b}_{K,j}|^{\frac{1}{2}}}, \tag{2.6}$$

with B_K representing the boundary contour in the keyframe containing the points $\mathbf{b}_{K,1}, \dots, \mathbf{b}_{K,|B_K|}$ and B_F being the corresponding boundary contour in the current frame containing the corresponding points $\mathbf{b}_{F,1}, \cdots, \mathbf{b}_{F,|B_F|}$. $\theta(\mathbf{c})$ is a normalization factor defined as follows:

$$\theta(\mathbf{c}) = \sum_{j=1}^{|B_K|} e^{-\ |\mathbf{c}-\mathbf{b}_{K,j}|^{\frac{1}{2}}}. \tag{2.7}$$

Each pixel $\mathbf{p}$ of the target region is then corrected by a bi-linear interpolation considering the coefficients of the four closest grid nodes.

2.4.2 Kawai *et al.*

Kawai *et al.* [59] analyse the neighboring area to estimate the variation of illumination in the target region. A grid $\mathscr{G}$ is defined on $\mathscr{N}_F \cup \mathscr{T}_F$ where each node is placed in the center of a patch. Initially, they assign the mean illumination variation of each patch in $\mathscr{N}_F$ to its corresponding node. The illumination variation in a pixel $\mathbf{p} \in \mathscr{N}_F$ is computed as described in the problem statement in equation (2.3). Then, for the target region, the illumination variation of each grid node is computed separately under the assumption that the change in brightness between two adjacent nodes is minimal. We define the function:

$$\psi_{\text{Kawai}} : \left\| \begin{array}{l} \mathscr{G} \subset \mathscr{N}_F \cup \mathscr{T}_F \rightarrow \mathbb{R} \\ \mathbf{c} \rightarrow \psi_{\text{Kawai}}(\mathbf{c}). \end{array} \right. \tag{2.8}$$

This function is obtained by minimizing the following global cost:

$$\min_{\psi_{\text{Kawai}}} \sum_{(\mathbf{c}_i,\mathbf{c}_j)\in P} (\psi_{\text{Kawai}}(\mathbf{c}_i) - \psi_{\text{Kawai}}(\mathbf{c}_j))^2, \tag{2.9}$$

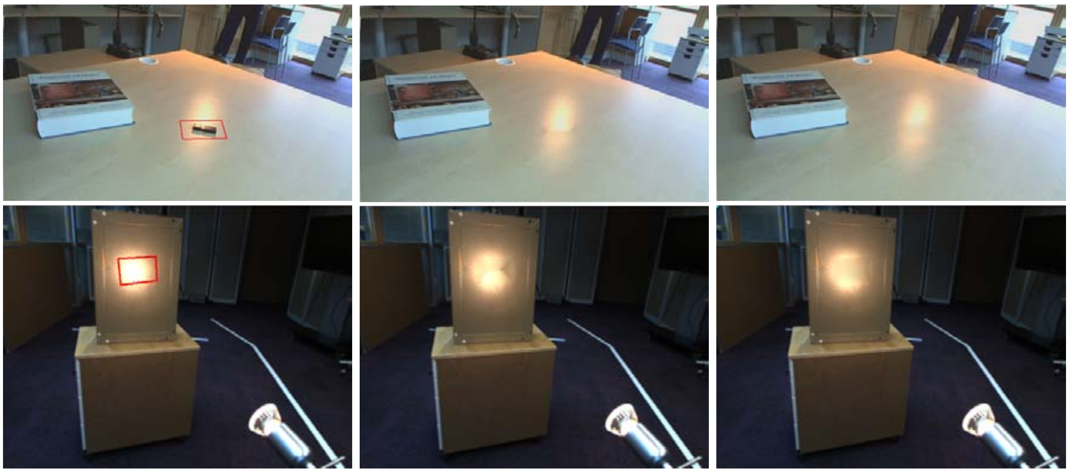
with, for $\mathbf{c}_i \in \mathscr{N}_F$:

$$\psi_{\text{Kawai}}(\mathbf{c}_i) = \frac{1}{|G_i|} \sum_{\mathbf{p}\in G_i} v_{\mathbf{p}}, \tag{2.10}$$

with $\mathbf{c}_i$ and $\mathbf{c}_j$ being the centers of two adjacent patches of the grid. P is a set of pairs of adjacent patches. G_i is the group of pixels in the patch centered around $\mathbf{c}_i$. Minimizing this cost allows one to retrieve the values of the grid nodes inside the target region $\mathscr{T}_F$. The coefficient of color variation $\psi_{\text{Kawai}}(\mathbf{p})$ for each pixel $\mathbf{p}$ inside the mask is then deduced by bi-linear interpolation.

2.4.3 Discussion

The two methods [49, 59] propose models that handle global image-level light changes well. However, only the smoothness property is considered by Kawai *et al.* (Property 1 in section 2). Herling *et al.*'s method [49] also respects the additivity property (property 5 in section 2). In other words, [59] uses a multiplicative model to update the illumination variation, [49] uses an additive model which is coherent with local illumination models. However, this is



(a) Original image (before DR) (b) Herling *et al.* (TVCG 2014) [49] (c) Kawai *et al.* (TVCG 2015) [59]

Fig. 2.4 Illustration of the limitations of previous methods. The images in column (a) represent the original image and the target area (outlined in red). The results of DR by the methods of Kawai *et al.* [59] and Herling *et al.* [49] are respectively shown in columns (b) and (c). These results are not visually convincing in both cases.

still insufficient in cases with specular surfaces and artificial lighting condition. Examples of DR show the limits of these methods in figure 2.4. Those demonstrate that even in basic scenarios including a specular planar surface under a point light source, state-of-the-art methods do not provide satisfying solutions.

2.5 Proposed Models and Methods

2.5.1 The Thin-Plate Spline

We propose two models based on our observations of the specularity's structural properties. Both models use the Thin-Plate Spline (TPS) as an interpolation function. It is a very suitable tool in this context because it enforces the smoothness constraint. Here, we briefly introduce the parameterization of the TPS. As inputs, we consider a set of l centers $\mathbf{c}_k \rightarrow u_k$ where $\mathbf{c}_k \in \mathbb{R}^2$ holds the coordinates of a center and $u_k \in \mathbb{R}$ is its corresponding unknown target value. We define the centers' coordinate matrix $\mathsf{C} = (\mathbf{c}_1 \cdots \mathbf{c}_l)$ and the centers' target vector $\mathbf{u}^\top = (u_1 \cdots u_l)$. The correspondence $(\mathbf{c}_1 \cdots \mathbf{c}_l) \rightarrow (u_1 \cdots u_l)$ represents the control points for the TPS. The TPS is a smooth function from $\mathbb{R}^2$ to $\mathbb{R}$ driven by these centers and given for any point $\mathbf{p} \in \mathbb{R}^2$ by:

$$\phi_{\text{tps}}(\mathbf{p};\mathbf{u}) = \mathbf{l}_\mathbf{p}^\top \ \mathsf{E}_\lambda \ \mathbf{u} \ , \tag{2.11}$$

where $\mathbf{l}_\mathbf{p}^\top = \Big(\rho\big(d^2(\mathbf{c}_1,\mathbf{p})\big)\cdots\rho\big(d^2(\mathbf{c}_l,\mathbf{p})\big)\Big)$ with $\rho(d) = d\log(d)$ being the TPS kernel for the squared distance. E_λ is the feature-driven parameterization matrix which incorporates an internal regularization weight $\lambda \in \mathbb{R}^+$ [12, 26]. λ controls the sensitivity of the interpolation function to fine variations. We set it to a small value for small-size specularities and a larger value for large-size specularities (λ can be set from 10^{-3} and up to 2.10^{-1}).

In practice, we have arbitrary positioned centers with unknown target values $\mathbf{u}$. So, given a set of m data points $\mathbf{q}_i \rightarrow v_i$, we estimate the optimal target values by solving:

$$\min_{\mathbf{u}} \sum_{i=1}^{m} (\phi_{\text{tps}}(\mathbf{q}_i;\mathbf{u}) - v_i)^2. \tag{2.12}$$

This forms a linear least squares problem, which we solve with a simple matrix pseudo-inverse.

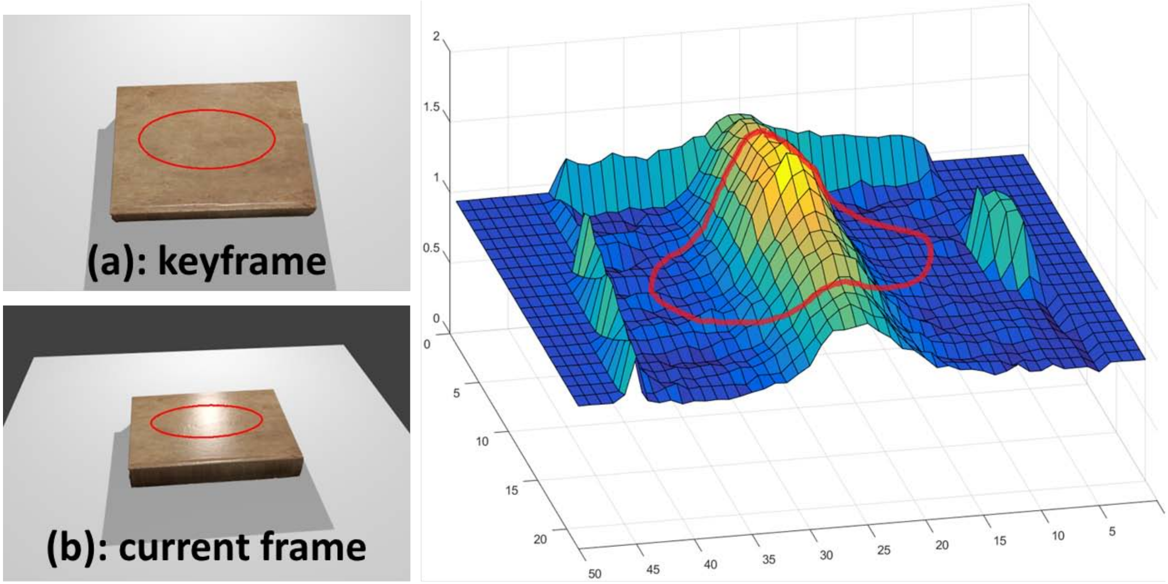


Fig. 2.5 The illumination variation in the image plane between the keyframe K and the current frame F is viewed as an elevation map. In this example, a specularity crosses the target region $\mathcal{T}_F$. This demonstrates the smoothness and continuity properties of a specularity.

2.5.2 Smooth Propagation Model

Description

The illumination variation can be viewed as a time-varying elevation map, as shown in figure 2.5. The base represents the pixel coordinates in the image and the height gives the variation's value. We propose a first model that only incorporates the smoothness property. We model the illumination variation by a TPS. This model makes few assumptions on the scene so it works for general case scenarios. However, it may generate poor results in some cases. We call it Smooth Propagation Model (SMP).

Estimation

We use the TPS to represent the function ψ_F that returns the illumination variation, for all pixels in $\mathcal{N}_F \cup \mathcal{T}_F$:

$$\psi_F : \left\| \begin{array}{l} \mathcal{O} \subset \mathbb{R}^2 \to \mathbb{R} \\ \mathbf{p} \to \begin{array}{l} v_{\mathbf{p}_K} = \chi_F(\mathbf{p}_F) - \chi_K(\mathbf{p}_K), \text{ if } \mathbf{p}_F \in \mathcal{N}_F \\ \phi_{\text{tps}}(\mathbf{p}_K;\mathbf{u}), \text{ if } \mathbf{p}_F = \eta_F(\mathbf{p}_K) \in \mathcal{T}_F \end{array} \end{array} \right. \tag{2.13}$$

We consider a uniformly distributed grid $\mathscr{G}_{\mathscr{F}} \subset \mathscr{N}_{\mathscr{F}} \cup \mathscr{T}_{\mathscr{F}}$. We set the grid so as to have l nodes, with l a perfect square. Using the parameterization of the TPS introduced in section 5.1, we consider the grid nodes as the centers $\mathbf{c}_k$ and the pixel intensity variations between the keyframe and the current frame as the target values u_k. The points in $\mathscr{N}_F$ are considered as the data points used to estimate $\mathbf{u}$. Using the estimation method from section 5.1, we obtain the TPS function ϕ_{tps}. The number of centers is chosen as $l = 100$, and the number of data points m depends on how many pixels we have in the neighboring region. In terms of computation, this method requires a least squares fit at every frame to solve (2.12). However, the matrix E_λ is constant, meaning that it can be precomputed from the keyframe only. In other words, solving for ψ_F requires solving minimization (2.12) with a simple multiplication between a constant matrix and the measured vector of variations $\mathbf{v}^\top = (v_1 \cdots v_m)$. In the RGB color space, we need to estimate separately three intensity differences for each pixel $\mathbf{p}_F$, E_λ being the same for the three color channels. A TPS was already used in [97] to model image-based light changes in the context of registration.

2.5.3 Constrained Propagation Model

Description

We extend SPM by considering more structural properties of specularities in the case of a scene with a single point light source. In this case, we constrain the new model by all the five properties from section 2. We call the second model Constrained Propagation Model (CPM).

Estimation

We integrate the constraint of the elliptical isocontours (property 3) by fixing a number s of intensity levels in the specular highlight. An isocontour is a set of pixels with the same intensity level (see figure 2.2 (b)). For each isocountour with intensity level h, we estimate the ellipse E by solving:

$$\min_{\mathbf{e}} \sum_{j=1}^{r} \Big(h - \chi^L\big(E(j)\big)\Big)^2, \tag{2.14}$$

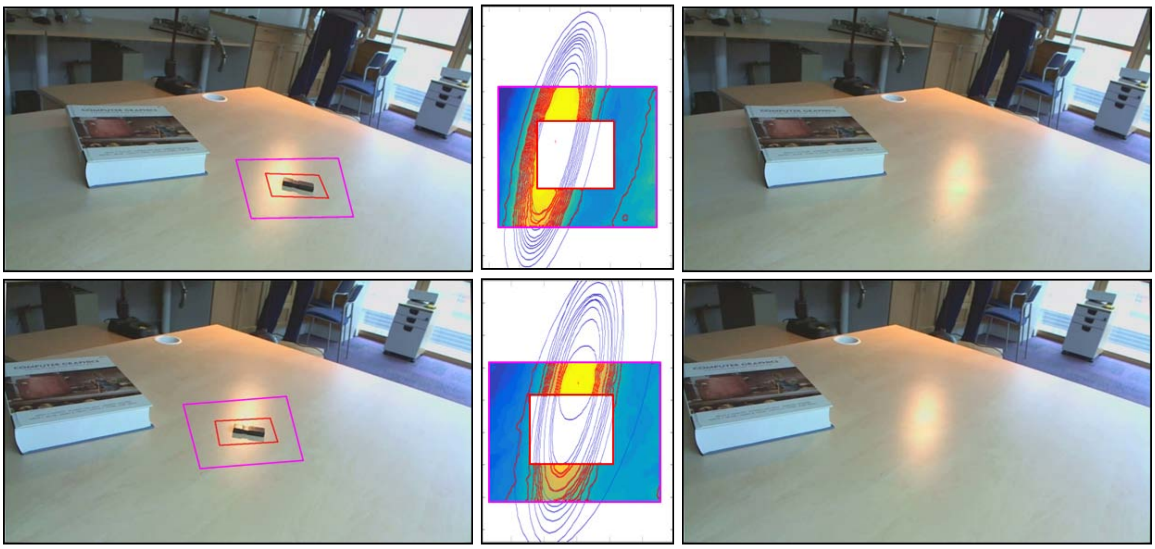
where χ^L returns the L color intensity values of a pixel in the Lab color space. The ellipse is represented by its five natural parameters $\mathbf{e}^\top = (o_x\ o_y\ a\ b\ w) \in \mathbb{R}^5$ with o_x and o_y as the center's coordinates, a as the semi-major axis, b as the semi-minor axis and w as the angle orienting the major axis. The ellipse is discretized in a group of points E of size $r = 100$ to evaluate the cost in (2.14), with $E(j) \in \mathbb{R}^2$ the j-th element in E. Further details on the fitting algorithm are given in section 5.3.2.2. The number of intensity levels s also represents the number of iso-contours considered, corresponding to the levels $h_{min} \cdots h_{max}$. Since the maximum intensity level h_{max} is constant, s is automatically adjusted depending on the minimum intensity level h_{min}. This threshold is fixed manually depending on the light exposure of the camera and light intensity. In particular, the brighter the light reflection, the higher the value of h_{min}.

Isocontour Detection

To evaluate the illumination variation, we convert the image to the *Lab* color space and consider only the L channel (Lightness). To reduce the computation cost, we only search for isocontours where a specularity is detected. To do so, we use a real-time algorithm for detecting specular reflections inspired from the methods in [82, 62]. To properly detect isocontours of intensity levels as ellipses, we begin by applying the Wiener filter [106] to segment the Lightness levels and reduce the noise generated by the roughness in the surface. Then we use a quantification histogram to segment the image into light intensity levels. The result is the brightness map. The detection of isocontours is carried out in this map. For a light level h, a point from the isocontour is detected when its corresponding intensity level is h and one of its neighboring points has an intensity level of *h-1*. Accordingly, we define s corresponding levels of intensity and detect their isocontour points.

Ellipse Fitting

Considering the brightness map obtained in the neighboring region $\mathscr{N}_F$, as shown in figure 2.6. The isocontour points that are outside $\mathscr{T}_F$ may be interpolated into ellipses extending within the target region. Our goal here is to estimate the extension of these isocontours assuming they have an elliptic shape. This is equivalent to minimizing the criterion in equation (2.14). To do so, we use the algorithm of Fitzgibbon *et al.* [29] for feature-based least squares fitting of ellipses. From a set of points from an isocontour, this method attempts to adjust the best ellipse which minimizes the algebraic distance. The result is used as an initialization. The parameters of the ellipse are fed to a simplex direct search algorithm to solve problem (2.14). This algorithm not only allows us to find all the points of the isocontour that are within the target area, but also the position of the ellipse's center. This information will be exploited to approximate the position of the brightest point. In this model we assume that the position of the brightest point is located at the center of the smallest detected isocontour. For an isocontour C, the difference of light variation for a color channel



(a) Original images (b) Ellipse fitting (c) Results of model 2

Fig. 2.6 Isocontour estimation on the brightness map. In (b), we use red represent the points used to estimate the ellipses and blue for the final estimation. The interpolated ellipses may extend inside the target region, and this allows us to propagate the specularity with high accuracy.

z is computed for points $\mathbf{p} \in C \cap \mathscr{T}_F$ as:

$$v^z_{\mathbf{p}} = \max_{\mathbf{q}\in C\cap\mathscr{N}_F} v^z_{\mathbf{q}}. \tag{2.15}$$

with C the set of points in the real isocontour and $v^z_{\mathbf{p}}$ the difference in intensity between $\mathscr{K}$ and $\mathscr{F}$ at pixel $\mathbf{p}$.

Consistency Filtering

In practice, due to estimation errors, there may be overlapping ellipses (see figure 2.7). So, we constrain the estimation of the ellipses by imposing property 4 from section 2, which is the monotonic decrease in intensity of the specular highlights. Based on our formulation of the problem, this property translates to the fact that *each ellipse of intensity level h_e should be totally inside any ellipse of intensity level $h < h_e$.* In order to respect this condition, we define a confidence coefficient for each estimated ellipse. The confidence coefficient for an estimated ellipse of intensity level h_e is determined with the function γ defined as:

$$\gamma(e) = \sum_{j=1,j\neq e}^{s} \delta(j,e), \tag{2.16}$$

where:

$$\delta(j,e) = \begin{cases} 1(E_j \subset E_e) \text{ if } h_j > h_e \\ 1(E_e \subset E_j) \text{ if } h_j < h_e \\ 0 \text{ otherwise.} \end{cases} \tag{2.17}$$

After the initialization with the feature-based fitting of ellipses, we compute the confidence coefficients of the different estimated contours and we retain the largest set of consistent ellipses (with the maximum confidence values). We refer to this operation by consistency filtering as in figure 2.7. We then refine the retrieved ellipses by solving problem (2.14).

Incorporating the Constraints to the TPS Returning to the estimation of the TPS, we add a fixed number of pixel coordinates as data points. They belong to the interpolated ellipses $E_1, \dots, E_s$ which are within the target region. Their respective target values are the differences in intensity of the corresponding isocontours.

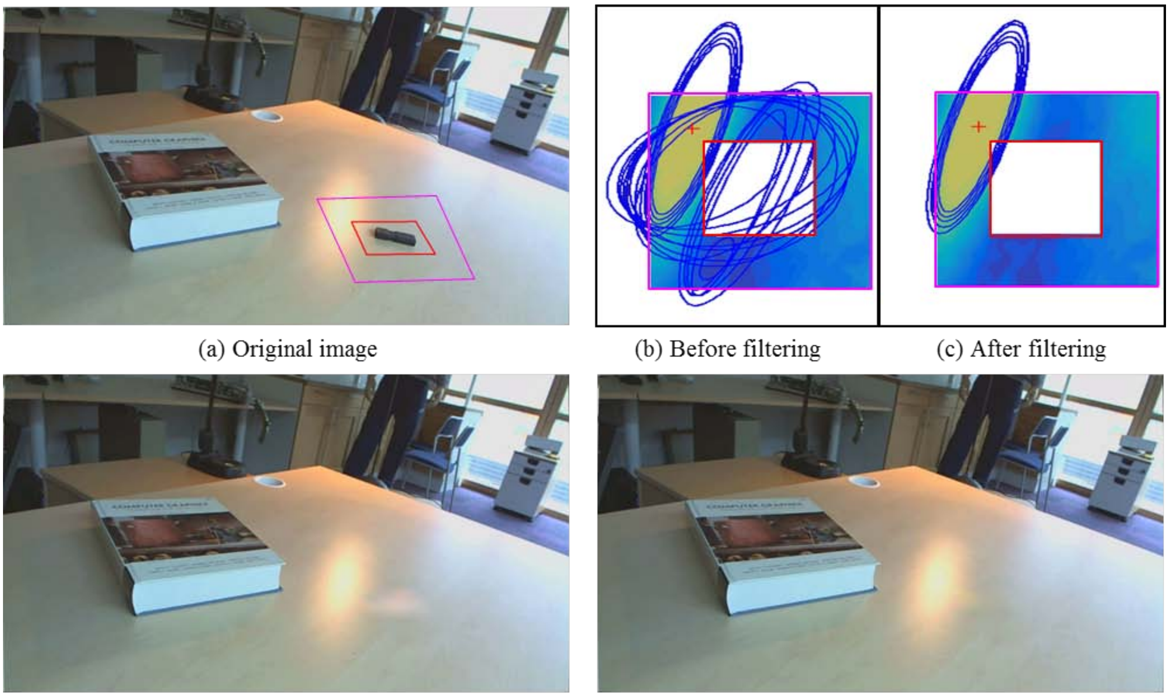
2.6 Results and Discussion

2.6.1 Datasets

The video sequences used for comparison in this section are divided into two categories. The first category includes two videos: a synthetic one (video 1) which was generated from the rendering software Blender-3D illustrated in figure 2.8, and a real one (video 2) illustrated in figure 2.9. The synthetic environment chosen in video 1 allows us to take full control of the different parameters in the scene (the light source's position, its intensity value, the object's material, camera orientation, etc). The reflection model is Phong's. For both videos, no undesired object is placed in the target region in order to let us compare the rendered specularity to the real one. The second category includes two real videos, video 3 (figure 2.10) and video 4 (figure 2.11), with undesired objects. These videos show the case of a planar surface crossed by a specularity.

2.6.2 Comparing Results

We can see that CPM achieves the best results for all videos in terms of specularity propagation. For the first two videos, CPM is the closest to the ground truth video. For the two other videos, its results are more visually coherent than for the other methods. So, this model shows an improvement of the rendering quality compared to SPM and state-of-the-art. This shows the relevance of incorporating the specularity's structural properties.



(a) Original image

(b) Before filtering

(c) After filtering

(d) Result before filtering

(e) Result after filtering

Fig. 2.7 Comparing the ellipse fitting results on the same video frame before and after filtering out the inconsistent ellipses. In (b), before filtering, many false estimations of ellipses are seen. This generates visual artifacts on the final result seen in (d). In (c), we were able to filter the false estimated ellipses according to their confidence coefficients as described in section 5.4.2. In (e), the final result of CPM is visually more convincing than the one in (d). The brightness maps in (b) and (c) are rectified to the keyframe image plane and zoomed for better visualization.

Even though it incorporates only the smoothness property, SPM still achieves better results than state-of-the-art. In fact, this demonstrates that the TPS is a well-adapted interpolation function for this problem. In particular, the method of Kawai *et al.* [59] works well with weak illumination variations when the brightest point is outside the target region. However, it clearly fails when a specularity enters the target region, with high intensity variations or with rich-texture surfaces. The method of Herling *et al.* [49] is very dependent on the boundary's shape. Therefore, the specularity rendering result is not visually convincing. It generates artifacts that seem unnatural. The interpolation of illumination variation in the previous methods does not respect the specularity's structure, which explains these results. Video 4 presents a very difficult case of rich texture surface (strong local variations of colors). In fact, the presence of such texture in the neighboring region along with the presence of the white color results in a non-smooth variation of pixel intensities in the three color channels. Since these pixels can be saturated in one or more of the RGB channels, the mean computed

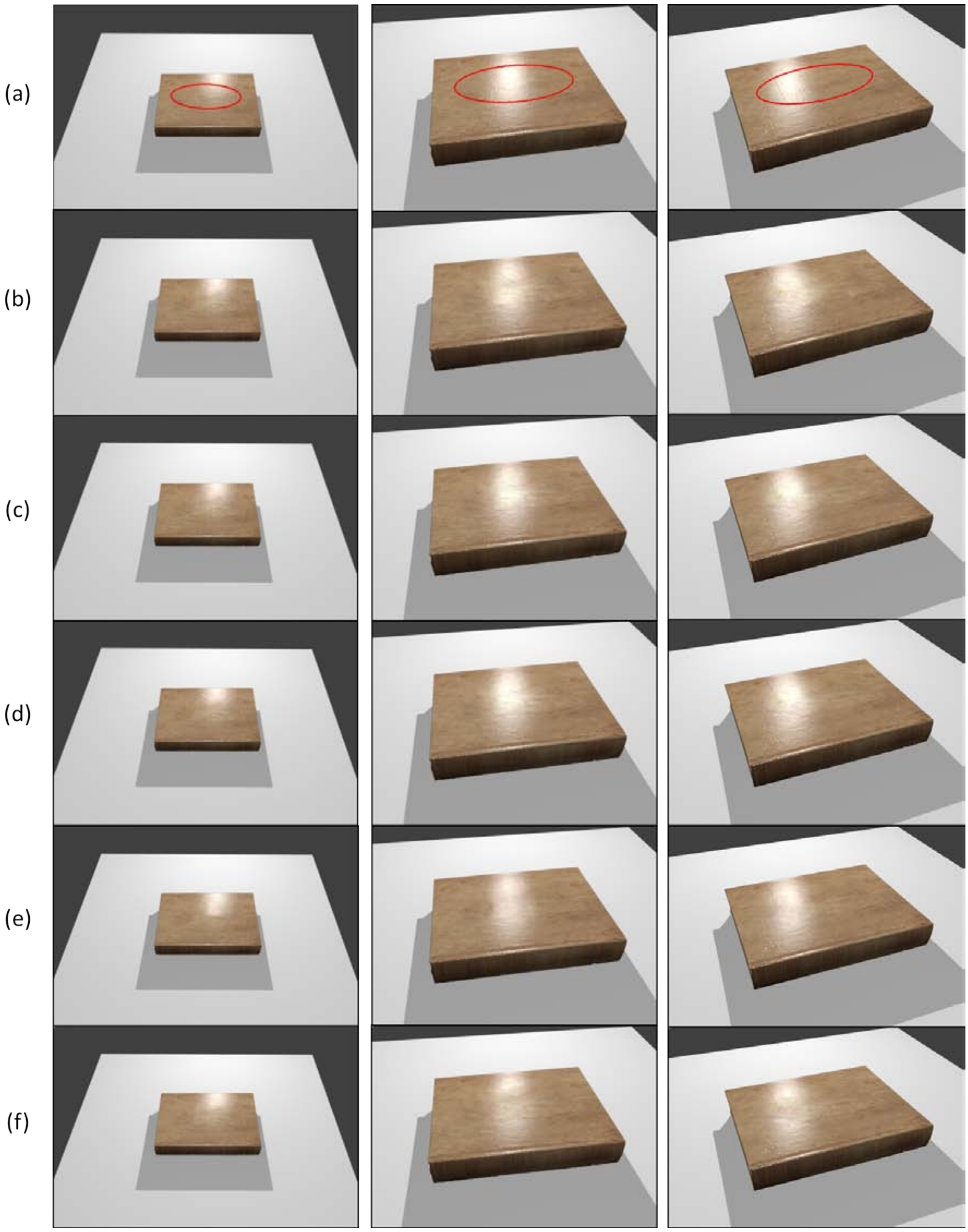


Fig. 2.8 Results of all the methods on frames 64, 71 and 130 of video 1. (a) corresponds to the original image with the target region in red. (b) SPM. (c) CPM before filtering inconsistent ellipses. (d) CPM after filtering inconsistent ellipses. (e) Method of Herling *et al.* [49]. (f) Method of Kawai *et al.* [59].

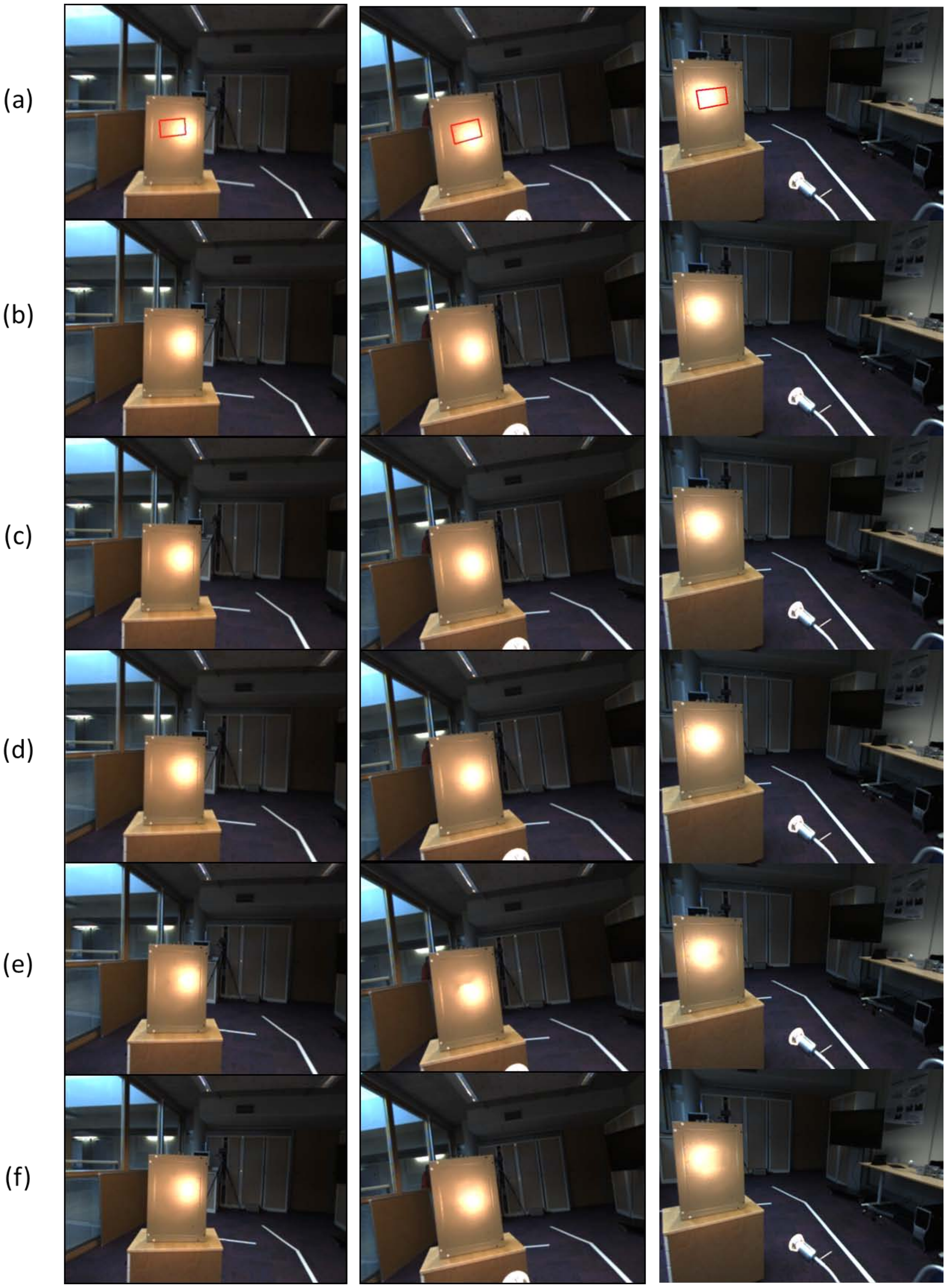


Fig. 2.9 Results of all the methods on frames 367, 541 and 1227 of video 2. (a) corresponds to the original image with the target region in red. (b) SPM. (c) CPM before filtering inconsistent ellipses. (d) CPM after filtering inconsistent ellipses. (e) Method of Herling *et al.* [49]. (f) Method of Kawai *et al.* [59].

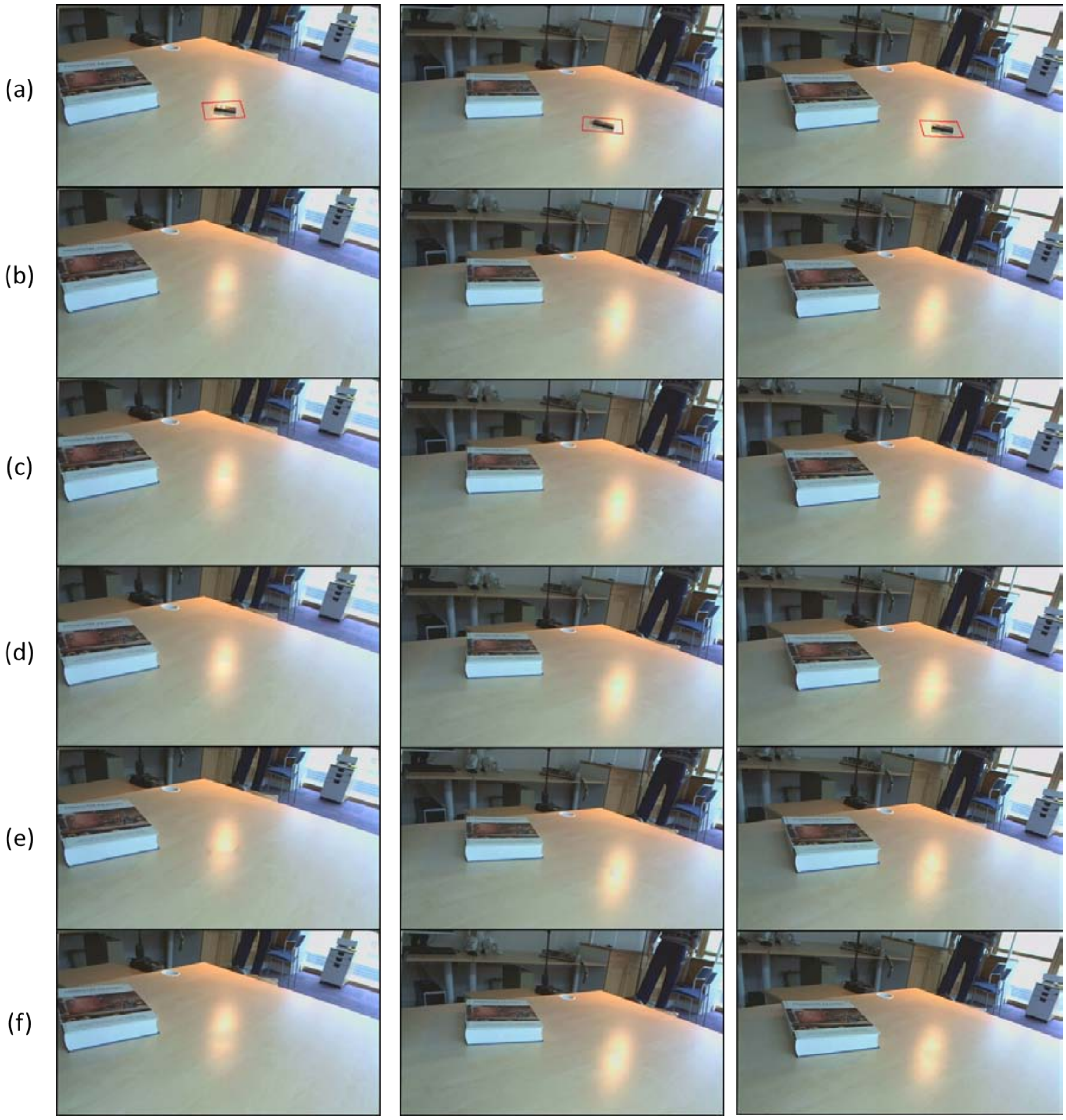


Fig. 2.10 Results of all the methods on frames 168, 202 and 328 of video 3. (a) corresponds to the original image with the target region in red. (b) SPM. (c) CPM before filtering inconsistent ellipses. (d) CPM after filtering inconsistent ellipses. (e) Method of Herling *et al.* [49]. (f) Method of Kawai *et al.* [59].

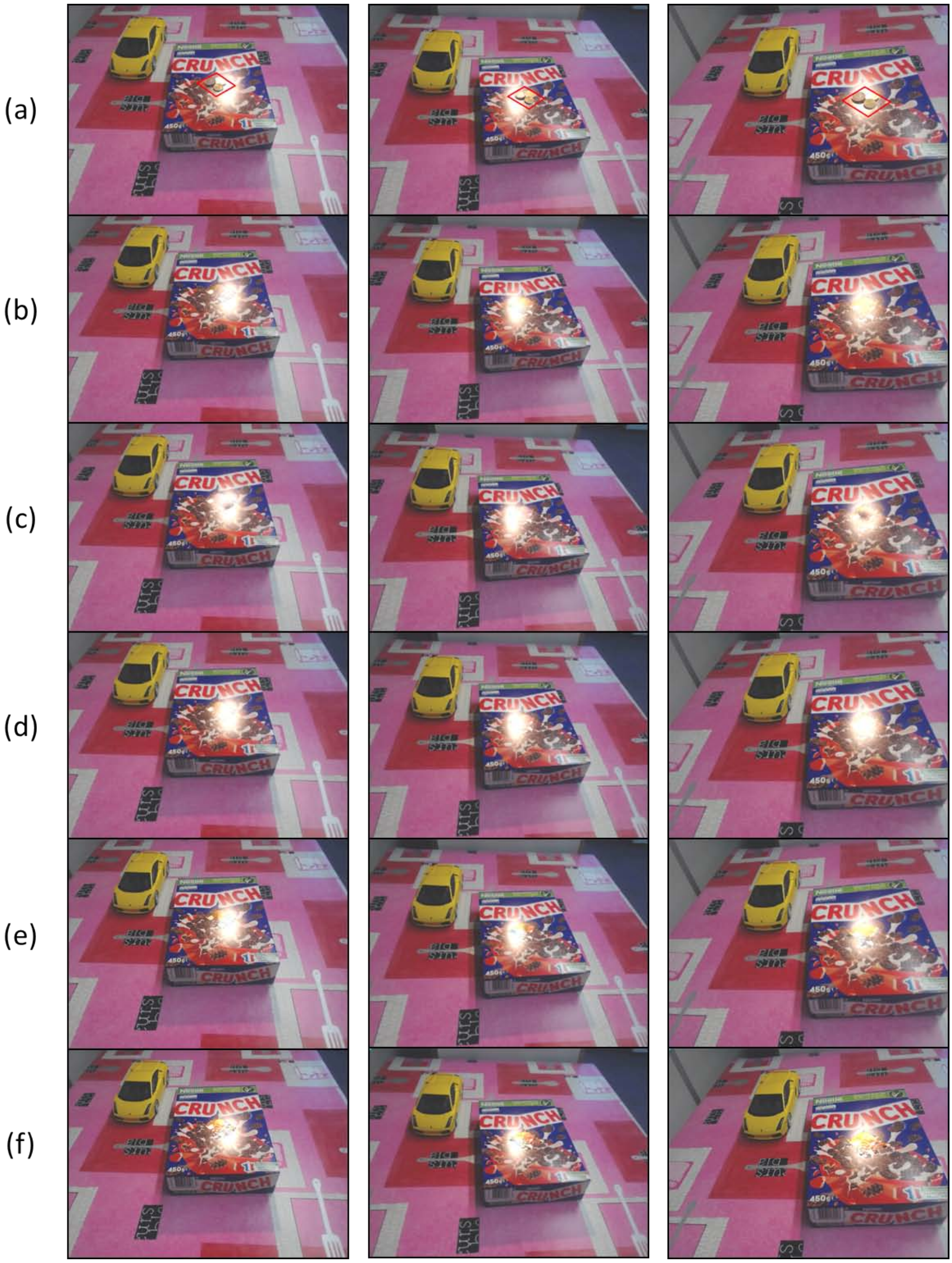
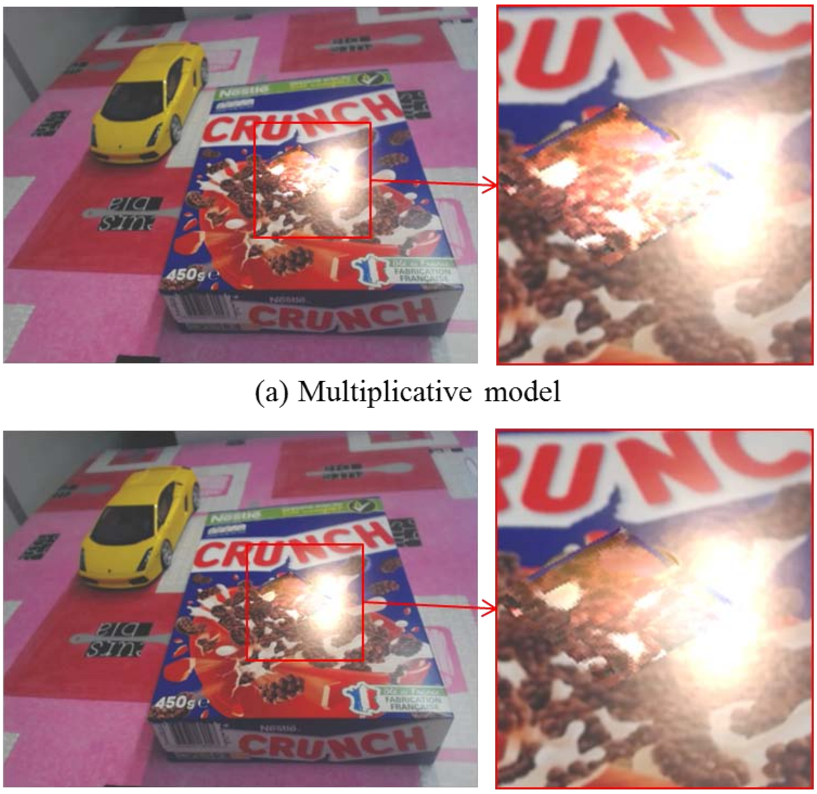


Fig. 2.11 Results of all methods on frames 637, 684 and 945 of video 4. (a) corresponds to the original image with the target region in red. (b) SPM. (c) CPM before filtering inconsistent ellipses. (d)CPM after filtering inconsistent ellipses. (e) Method of Herling *et al.* [49]. (f) Method of Kawai *et al.* [59].



(a) Multiplicative model

(b) Additive model

Fig. 2.12 Comparing the results of our second model CPM using an additive model versus a multiplicative model for computing the illumination variation coefficients.

variation is usually insufficient to reproduce the specularity for SPM and previous methods. However, CPM can overcome this issue by imposing additional constraints. The fact that we consider the maximum difference of intensity for the isocontour points allows us to avoid these extreme cases. Although some artifacts can still occur, our model gives the best results by far, compared to previous methods (see figure 2.11).

2.6.3 Additive Versus Multiplicative Model

As mentioned in section 4, the previous methods express differently the illumination variation coefficients. Herling *et al.* [49] use an additive model while Kawai *et al.* [59] use a multiplicative model. In our proposed methods, our choice was guided by the local illumination models. This choice is also supported by the results on real data. In figure 2.12, we compare the results of CPM associated for the two different possibilities. This shows that the pixel colors are altered when using a multiplicative model (they become overly bright). This is particularly strong for rich texture surfaces. This is also confirmed by the results of

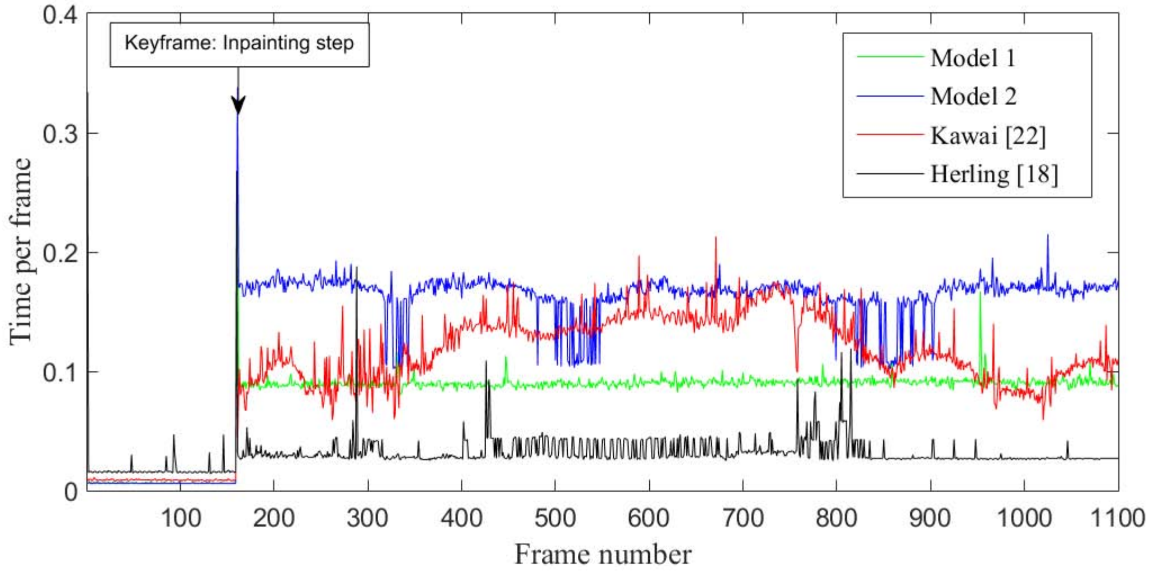


Fig. 2.13 The computation time per frame in seconds for the proposed models and previous models on video 2.

[49] (using the additive model) which are more convincing than the ones of [59] (using the multiplicative model) on video 4.

2.6.4 Computation Time

In figure 2.13, we present the computation time for each frame of video 2 using the proposed models as well as the state-of-the-art methods. We ran these tests on an Intel i7 processor with a 2.70 GHz frequency. Video 2 has a frame size of 640x480 pixels. The inpainting step is launched at frame 160, which explains the peak in computation time for all models. SPM has a stable computation time at around 70 ms per frame which corresponds to a performance of 14 frames per second. CPM has a mean computation time of 160 ms per frame which corresponds to a performance of about 7 frames per second. Many optimizations could be done, including the parallelization of the algorithm and an optimized choice of the parameters. However, we can already state that the proposed models are well adapted for real-time applications.

2.6.5 Discussion and Limitations

Our proposed model SPM imposes the smoothness property of a specularity which is true for any type of specular surfaces. This property is observed on most types of surfaces and it is also imposed by state-of-the-art methods. However, our model uses a more adapted mathematical formalism based on the TPS which generates more plausible results. Our proposed CPM incorporates further properties based on the specularity's shape. However, the property of the uniqueness of the brightest point is particularly valid for a point light source. Under these circumstances, CPM outperforms by far previous methods as shown on the experimental results. The choice between the first and second models can be decided depending on the specifications of the observed scene. The methods presented in this paper consider a single plane in the target region and therefore compute a single homography corresponding to this plane. Kawai *et al.* [59] propose to consider multiple homographies in order to handle multiple planes. Similarly, our estimation method can also be extended by considering the geometry of the surface if known.

Our models along with state-of-the-art do not handle a rough surface because in this case the property of smoothness is no longer valid. Another specific case, which is not considered by our models as well as state-of-the-art, is when a small specularity fully enters the target region. For this case, a full prediction of the illumination in the scene is needed. This implies the estimation of material properties as well as light source configuration. Chapter 3 addresses this problem as a continuity to our work.

2.7 Conclusion

In this chapter, we dealt with the illumination variation problem in the context of Diminished Reality. This complex problem can be transformed into a specularity propagation problem. From multiple empirical observations, we proposed a list of structural properties of a specularity. We then proposed two models that embed these properties to estimate the illumination variation. Our first model is generic, while our second model is more adapted to curvy surfaces with single point light sources. Our experimental results show the relevance of our approach compared to previous work which do not embed these properties. Particularly, the results of our second model CPM show a substantial improvement in rendering results with respect to the specularity's spatial structure compared to state-of-the-art. In the next chapter, we evaluate an approach to estimate scene's parameters with respect to the local illumination models which allows a more general solution to specularity propagation in Diminished Reality as well as Augmented Reality.

Chapter 3

Illumination Model Inversion in Mixed Reality

The work reported in this chapter was published in IEEE Eurographics [42].

3.1 Introduction

The specularity propagation in Diminished Reality can also be solved using local illumination models which generate all the image's components including the specular highlights. However, the parameters of the scene exploited by these models are often unknown. The problem of retrieving these parameters is the local illumination inversion problem. In this chapter, we address this problem in the context of both Augmented and Diminished Reality applications. It is known that global illumination is an accurate model, because it considers both direct and indirect illumination and therefore can generate all levels of light reflections. However, all previous approaches ignore indirect illumination because it makes the problem intractable. Some methods represent illumination by alternative models. Debevec *et al* [24] model the incident light by a spherical environment map obtained from a light probe. In their approach, a light probe must be installed in the scene and must be clearly visible in the images. Along with the light probe constraint, this method assumes distant lighting. The environment map is only valid for a limited region and is therefore not a good approximation for points far off the light probe. Some works attack the problem using specularities [40, 54, 84]. We refer to them as *specularity modeling* methods. Jachnik *et al* [54] estimate the environment map by capturing the surface light-field from multiple images. The light-field is then transformed into hemispheres representing the diffuse and specular components of the surface. The specular component is projected to the surface to predict its value at each viewpoint.

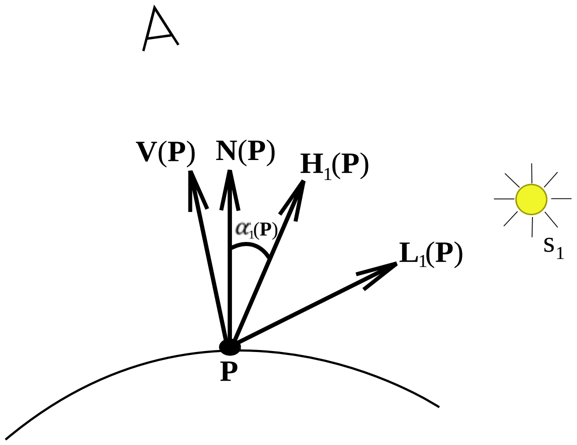


Fig. 3.1 Notation used to formulate the reflectance models at a 3D point **P**.

However, this solution only works for flat surfaces and is limited to a specific region in the scene. More recently, Morgand *et al* [84] have modeled the specularities using a 3D quadric that is reconstructed from at least three images. The projection of the quadric by the camera allows them to predict the specular component in new viewpoints. Our method presented in chapter 2 falls into the same category which consists of a 2D representation of specularities in the context of DR [40]. The specularity modeling methods [40, 54, 84] obtain convincing results for real-time AR and DR and show the importance of the specularity as a visual cue. However, they model an abstract object (hemisphere [54], quadric [84], ellipse [40]) that is tied down to a specific material under specific illumination conditions. In other words, they merge the lighting and material properties into a joint element. Consequently, the user cannot edit the scene's physical properties in their representation.

An alternative to specularity modeling is *inverse local reflectance* approaches which unambiguously separate the light source position and intensity, the material parameters, the camera parameters and the scene's geometry. Contrary to global illumination, local illumination only considers direct reflection. In local reflectance inversion, parameterized BRDF models (Bi-directional Reflectance Distribution Function) such as Phong's [89] are estimated to fit the real scene images. This allows one to meet the real-time reconstruction constraint and enhance rendering flexibility and precision. This also makes it possible to render shadows and specularities from new viewpoints and to estimate separately the properties of materials and lighting. Assuming that the camera pose and geometry of the scene are known, as in [40, 54, 84], it is however still tremendously challenging to estimate the parameters of the reflectance models reliably. For each surface and each light source, the parameters to estimate are the light source's position and intensity and the surface's material reflectance and roughness.

Local reflectance inversion has received several computational solutions for one or several images. However, the well-posedness of the general problem was rarely studied in the previous approaches. Depending on the input data, it is not clear whether the solution would be unique and well-constrained, or not. This is a key point that needs clarification for the sake of realism in AR and DR. The answer clearly depends on the number of input images but also on the type of visual cues. Some works use a single image [11, 44] and some use multiple [80, 87, 107]. The recent approaches of specularity modeling [54, 84, 40] show that specular highlights form perhaps the most important visual cues to solve local reflectance inversion. The study by Morgand *et al* [84, 81] has shown that one can associate a virtual camera, and thus a unique virtual viewpoint, with each specular highlight. The consequence is that several specular highlights, even when taken from a single image, will exert complementary constraints on the sought local reflectance parameters. Also, Rammamorthi *et al* [92] as well as Yu and Debevec [108] stated that the specular highlight configuration in the input images is a key factor in the well-posedness of the problem. However, they did not bring a formal conclusion on the minimal required configuration to solve the problem. We now introduce three key definitions needed to characterize the input data in local reflectance inversion problems.

Definition 1 *Specular highlight. A specular highlight is a connected region of the image where the observed intensity is predominantly due to the direct reflection of the incoming light. It is characterized by a single maximum of intensity. The specular highlight may also include a low intensity diffuse component.*

Definition 2 *Single-spot method. A method for local reflectance inversion is called single-spot when it uses a single specular highlight as input.*

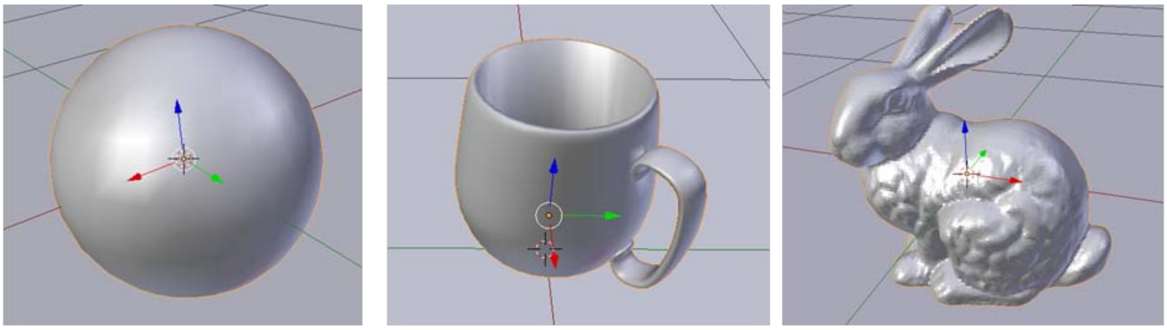
Definition 3 *Multi-spot method. A method for local reflectance inversion is called multi-spot when it uses multiple specular highlights as input. These may come from one or several images but from only one light source.*

While a multi-spot approach may provide richer data, it comes at a price. Indeed, two highlights strengthen the setting only if they are associated to the same light source. A true multi-spot approach will thus require a specularity matching or tracking engine in order to assign each observed specularity to its corresponding light source. It will also require that the number of light sources be known or estimated by some mechanism. In contrast, all this is not needed in the single-spot approach. Motivated by this observation, we propose to assess to which extent the local reflectance inversion problem is well-constrained, considering as visual cue a single specular highlight. In other words, we address the following question: *can we* *invert a local reflectance model with the single-spot approach knowing the scene geometry and camera pose?* We attempt to answer this question empirically for three reflectance models, namely Blinn-Phong, simplified Torrance-Sparrow and Ward for isotropic surfaces. In this chapter, *we exhaustively investigate the correlation between the input data and the efficiency of local reflectance inversion.* We consider a region of the input image associated with a single specular highlight as input data. The parameters of the reflectance models are retrieved by minimizing the photometric cost expressed by the least squares difference between the input and the predicted image.

3.2 State-of-the-Art

Inverse rendering is a well-studied problem in Computer Graphics and Computer Vision. It aims to recover scene parameters from a single or multiple images. The scene parameters include geometry, lighting, reflectance and camera properties. Since many approaches in Computer Vision robustly recover surface geometry [53] and camera parameters [99], we only discuss the approaches that recover lighting and reflectance. Patow *et al* [88] classify the inverse rendering problems into three categories: inverse lighting, inverse reflectometry and combined inverse lighting and inverse reflectometry problems. In the inverse lighting problem, the goal is to recover the properties of light sources in the scene with known surface parameters. The unknown properties include the light source positions as well as their intensities in some color space. In this category, two types of assumption are commonly used for the lighting. Many assume distant lighting and therefore only recover the direction and intensity of light sources [1, 66, 76, 94]. This includes the recent methods that use Deep Learning to estimate the illumination map [31, 34]. These approaches are relevant for outdoor scenes. However, in indoor scenes, the estimated illumination is only valid in a specific and limited region. Others assume that light sources are points in space [13, 16, 27, 91, 105]. So, a limited number of point light sources are estimated in terms of 3D position and intensities. In the inverse reflectometry problem, the lighting parameters are assumed known while surface reflectance (BRDF) is unknown. To solve this problem, most of the approaches [25, 93, 11, 108, 74] assume an homogeneous BRDF with constant albedo, and that the surface's reflectance can be modeled by an analytical BRDF model such as Ward's. Other approaches consider a spatially-varying BRDF but with different additional assumptions. In [4, 109], the materials are considered to be Lambertian and therefore only the diffuse component is estimated. The intensity associated to every surface point is therefore constant regardless of the viewing direction. In [52, 35], the surface's BRDF is described as a convex combination of a small number of fundamental materials.

In a vast majority of their applications, AR and DR impose the third category of the inverse rendering problem: combined inverse lighting and reflectometry. In these applications, geometry and camera parameters can be acquired with sufficient accuracy using RGB-D cameras and the SLAM approach. We classify existing approaches based on the BRDF model they use. Mercier *et al* [80] propose a complete environment for reconstructing an object from a set of images using a modified version of the Phong model. Xu *et al* [107] use two images acquired by a stereo camera to fit the Blinn-Phong model. Hara *et al* [44] use a polarizer to separate the diffuse and specular components from the image. They then estimate separately from each component the different parameters of the simplified Torrance-Sparrow model [102]. Finally, Ramamoorthi *et al* [92] propose a signal-processing framework that integrates both the Blinn-Phong and Ward models. So, different reflectance models are tested in the literature but with no assessment on which one is best for the local reflectance inversion problem and under which conditions.



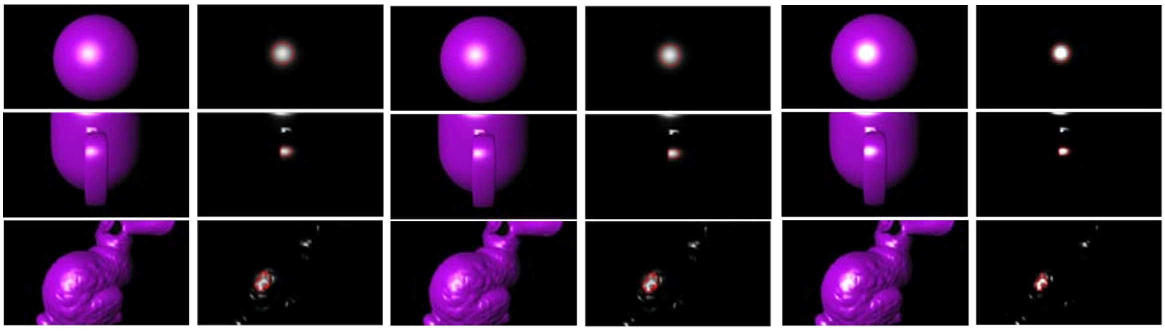
(a) Smooth convex object

(b) Smooth non-convex object

(c) Rough non smooth object

Fig. 3.2 The meshes of the 3D objects used for our image dataset.

In the context of AR and DR applications, we are interested in determining the minimal input data needed to solve the inverse rendering problem and quantifying the confidence level of the estimated parameters. This allows us to optimize resources and determine a well-posedness criterion. In the literature, the problem has been addressed using both single and multiple image approaches. However, in this specific problem a single image of a complex scene could include more visual cues than multiple images of a simple scene. In order to evaluate the well-posedness of the inverse local reflectance problem we therefore review the visual cues present in the input data. Some methods use shadows [94], some use specularities [54, 84, 40] and some use both [91]. Although shadows could include valuable information on the light source position, they do not contain information about the light intensity nor the surface parameters. In fact, they are only used in the inverse lighting problem. However, specularities provide information on all the reflectance model's



(a) Blinn-Phong (BP) (b) Torrance-Simplified (TS) (c) Ward-Isotropic (WI)

Fig. 3.3 A sample of images of the synthetic dataset under different camera poses and models of local reflectance. We consider a total of 81 images (3 reflectance models $\times$ 3 angles $\times$ 3 distances $\times$ 3 objects). The specular highlights extracted from the image are shown on the right. The region Ω containing the specular highlight is in red.

parameters. Morgand *et al* [84] provided the insight that each specularity can be associated with a virtual camera and therefore represents a key visual element on its own. From this observation, the well-posedness of the problem can be assessed in terms of the number of visible specular highlights.

As motivated, we assess the well-posedness of the inverse local reflectance inversion problem using a single specular highlight. The specular highlight configuration has already been related before to the well-posedness of the problem. Yu and Debevec [108] stated that to obtain an obvious global minimum for the inverse reflectometry problem, the radiance image should cover an area that has specular highlights as well as some areas with a very low specular component. Also, based on their signal processing framework, Ramamoorthi *et al* [92] noted that strong specular highlights are necessary for the well-conditioning of inverse lighting while soft specular highlights are needed for inverse reflectometry. These observations have not been experimentally qualified and a precise specular highlight configuration was not suggested in their work. In contrast, we propose to thoroughly investigate this open question. We define an experimental setup that assesses the well-posedness of the problem when taking into consideration a single specular highlight. We treated several representative scenarios to ensure completeness to our investigation.

3.3 Reflectance Models

We use the point light model, whose parameters are its 3D position and intensity in RGB[1].

[1] The RGB channel representation constitutes a commonly accepted approximation for the light intensity spectrum.

3.3.1 Notation

Scalars are in lowercase italics (x), vectors are in bold upright ($\mathbf{V}$) and functions similarly, depending on their returned parameter. We use $\odot$ to represent the element-wise multiplication between two vectors ($\mathbf{u} = \mathbf{v}_1 \odot \mathbf{v}_2$) and $\cdot$ to represent the scalar product ($x = \mathbf{v}_1 \cdot \mathbf{v}_2$). Figure 3.1 shows the basic elements we need for the reflectance models. For a surface point $\mathbf{P}$, we have $\mathbf{V}(\mathbf{P})$ the viewing direction, $\mathbf{N}(\mathbf{P})$ the surface normal and $\mathbf{L}_n(\mathbf{P})$ the direction to the light source s_n with $n \in [1, N]$, N being the number of light source. The halfway vector is $\mathbf{H}_n(\mathbf{P}) = \frac{\mathbf{L}_n(\mathbf{P})+\mathbf{V}(\mathbf{P})}{\|\mathbf{L}_n(\mathbf{P})+\mathbf{V}(\mathbf{P})\|}$. We refer to the angle between $\mathbf{H}_n(\mathbf{P})$ and $\mathbf{N}(\mathbf{P})$ as $\alpha_n(\mathbf{P})$. We consider $\mathbf{I}(\mathbf{P}) \in [0,1]^3$, the intensity of a surface point $\mathbf{P}$ in the RGB color space. We straightforwardly obtain the correspondence between image pixels and surface points using ray tracing knowing the 3D model of the surface and the camera pose.

3.3.2 Standard Models

A reflectance model is the sum of three components: ambient, diffuse and specular. The pixel intensity $\mathbf{I}(\mathbf{P}) \in [0,1]^3$ of a surface point $\mathbf{P}$ is modeled as:

$$\mathbf{I}(\mathbf{P}, \mathbf{k}_a, \mathbf{k}_d, \mathbf{k}_s, m, \mathbf{i}_a, \mathbf{i}_1 \ldots \mathbf{i}_N) = \mathbf{k}_a(\mathbf{P}) \odot \mathbf{i}_a + \sum_{n=1}^{N} \mathbf{N}(\mathbf{P}) \cdot \mathbf{L}_n(\mathbf{P}) \mathbf{k}_d(\mathbf{P}) \odot \mathbf{i}_n + \sum_{n=1}^{N} \mathbf{J}_s(\mathbf{P}, \mathbf{k}_s \odot \mathbf{i}_n, m, \mathbf{L}_n, \mathbf{i}_n). \tag{3.1}$$

The term $\mathbf{k}_a(\mathbf{P}) \odot \mathbf{i}_a$ represents the ambient component where $\mathbf{i}_a \in [0,1]^3$ is the ambient light intensity in RGB and $\mathbf{k}_a(\mathbf{P})$ the ambient reflectance coefficient at $\mathbf{P}$. This term approximates the effect of indirect lighting. The terms $\mathbf{N}(\mathbf{P}) \cdot \mathbf{L}_n(\mathbf{P}) \mathbf{k}_d(\mathbf{P}) \odot \mathbf{i}_n$ and $\mathbf{J}_s(\mathbf{P}, \mathbf{k}_s \odot \mathbf{i}_n, m, \mathbf{L}_n, \mathbf{i}_n)$ represent the contribution of the light source s_n to the diffuse and the specular components respectively, with $\mathbf{i}_n$ the intensity of the light source s_n in RGB, $\mathbf{L}_n$ its 3D position and $\mathbf{k}_d(\mathbf{P})$ and $\mathbf{k}_s(\mathbf{P})$ the diffuse and specular reflectance coefficients respectively. We compare three models used in the literature in the context of local reflectance inversion: Blinn-Phong [10] which we refer to as BP, a simplified version of Torrance-Sparrow [102, 85] which we refer to as TS and Ward [103] for Isotropic surfaces which we refer to as WI. These models follow equation (3.1). They differ solely by their specular component $\mathbf{J}_s(\mathbf{P}, \mathbf{k}_s \odot \mathbf{i}_n, m, \mathbf{L}_n, \mathbf{i}_n)$:

$$\mathbf{J}_s(\mathbf{P}, \mathbf{K}_s, m, \mathbf{L}_n, \mathbf{i}_n) = \begin{cases} \mathbf{K}_s \left(\mathbf{H}_n(\mathbf{P}) \cdot \mathbf{N}(\mathbf{P})\right)^m & \text{(BP)} \\ \frac{\mathbf{K}_s}{\mathbf{N}(\mathbf{P}) \cdot \mathbf{V}(\mathbf{P})} \exp\left(-\frac{\alpha^2(\mathbf{P})}{2m^2}\right) & \text{(TS)} \\ \frac{\mathbf{K}_s}{4m^2\sqrt{(\mathbf{N}(\mathbf{P}) \cdot \mathbf{V}(\mathbf{P}))(\mathbf{N}(\mathbf{P}) \cdot \mathbf{L}_n(\mathbf{P}))}} \exp\left(-\frac{\tan(\alpha(\mathbf{P}))^2}{2m^2}\right) & \text{(WI)} \end{cases} \tag{3.2}$$

3.3.3 Hypotheses

Since we consider a single-spot approach, we assume that *(i) a single light source* s_1 *contributes to the observed specularity*. This assumption holds even in the presence of multiple light sources because, even indoor, it is unlikely to have multiple light sources with the same direction. To reduce the complexity of the problem we also consider that *(ii) the object's surface has a constant albedo and roughness*. In real-world scenarios, texture-less surfaces made of a homogeneous material are largely available, especially indoor. These assumptions allow us to focus our investigation on the correlation between the input data and the well-posedness of the problem. It constitutes a basis that can be extended afterwards for studying more complex illumination configurations and isotropic or glossy materials.

3.3.4 Reduced Models

The first hypothesis allows us to retain only the contribution of the first light source to the specular and diffuse components. This means that the diffuse and specular components corresponding to secondary light sources are constant and can be absorbed by the ambient term. The second hypothesis implies that $\mathbf{k}_a$, $\mathbf{k}_d$ and $\mathbf{k}_s$ become independent of $\mathbf{P}$. Therefore, the reduced model $\mathbf{I}_e$ is expressed as:

$$\mathbf{I}_e(\mathbf{P},\mathbf{S}_1,\mathbf{K}_a,\mathbf{K}_d,\mathbf{K}_s,m) = \mathbf{K}_a + \mathbf{N}(\mathbf{P}) \cdot \mathbf{L}_1(\mathbf{P})\mathbf{K}_d + \mathbf{J}_s(\mathbf{P},\mathbf{K}_s,m,\mathbf{L}_1), \tag{3.3}$$

where $\mathbf{K}_a = \mathbf{k}_a(\mathbf{P}) \odot \mathbf{i}_a + \sum_{n=2}^{N} \mathbf{N}(\mathbf{P}) \cdot \mathbf{L}_n(\mathbf{P})\mathbf{k}_d(\mathbf{P}) \odot \mathbf{i}_n$, $\mathbf{K}_d = \mathbf{k}_d \odot \mathbf{i}_1$ and $\mathbf{K}_s = \mathbf{k}_s \odot \mathbf{i}_1$. Since we consider RGB images, all these parameters are in three dimensions. To sum up, we estimate:

- $\mathbf{K}_a \in [0,1]^3$: the constant representing the ambient component and the contribution of secondary light source to the diffuse component.
- $\mathbf{K}_d \in [0,1]^3$: the coefficients of the diffuse reflectance for the first light source.
- $\mathbf{K}_s \in [0,1]^3$: the coefficients of the specular reflectance.
- $\mathbf{S}_1 \in \mathbb{R}^3$: the position of the first light source.
- $m \in \mathbb{R}_+$: the roughness of the object's surface.

3.4 Proposed Methodology and Inversion Approach

The input image is written $\mathbf{I}_r$ and named original image. We denote its specular component, obtained by diffuse-specular decomposition or by direct rendering, as $\mathbf{I}'_r$. Details on obtaining this image are reported in the next paragraph.

3.4.1 Specular Highlight Region

An image region Ω is extracted as the reference data. This region corresponds to the largest specular highlight in the original image. To extract it, we consider the specular component $\mathbf{I}'_r$ of the image and segment the specular highlights by setting a threshold $\varepsilon = \theta I_{max}$ that is applied to the three RGB channels. We then obtain several connected regions. In our experiments, the maximum channel intensity is $I_{max} = 1$ and the minimum intensity is $I_{min} = 0$. The value of θ is set to respect the rule of a unique maximum intensity in the region. In accordance with the observation of Yu and Debevec [108], this region consists of a strong specular highlight plus an area with low intensity diffuse component. For synthetic images, the largest region is retained as the region of interest. For real images, we manually select the specular region depending on the object of interest. In figure 3.3, the specular highlight region considered in each of the examples is shown in red on the specular component images.

3.4.2 Scenarios

In order to assess the correlation between the input data and the well-posednes of this problem, we propose different test scenarios. Three main scenarios are investigated as shown in table 3.1. In scenario **P1**, only the specular component $\mathbf{I}'_r$ is used as input. In this case, we also investigated the correlation between the convergence of the algorithm and each estimated parameter in the specular component. In scenario **P2**, the original images including the three components are used (without any specular-diffuse decomposition). In scenario **P3**, we use the results from **P1** as initialization and apply the optimization algorithm on the three component images. For each scenario, we also compared the results of combining or separating the color channels in the optimization process.

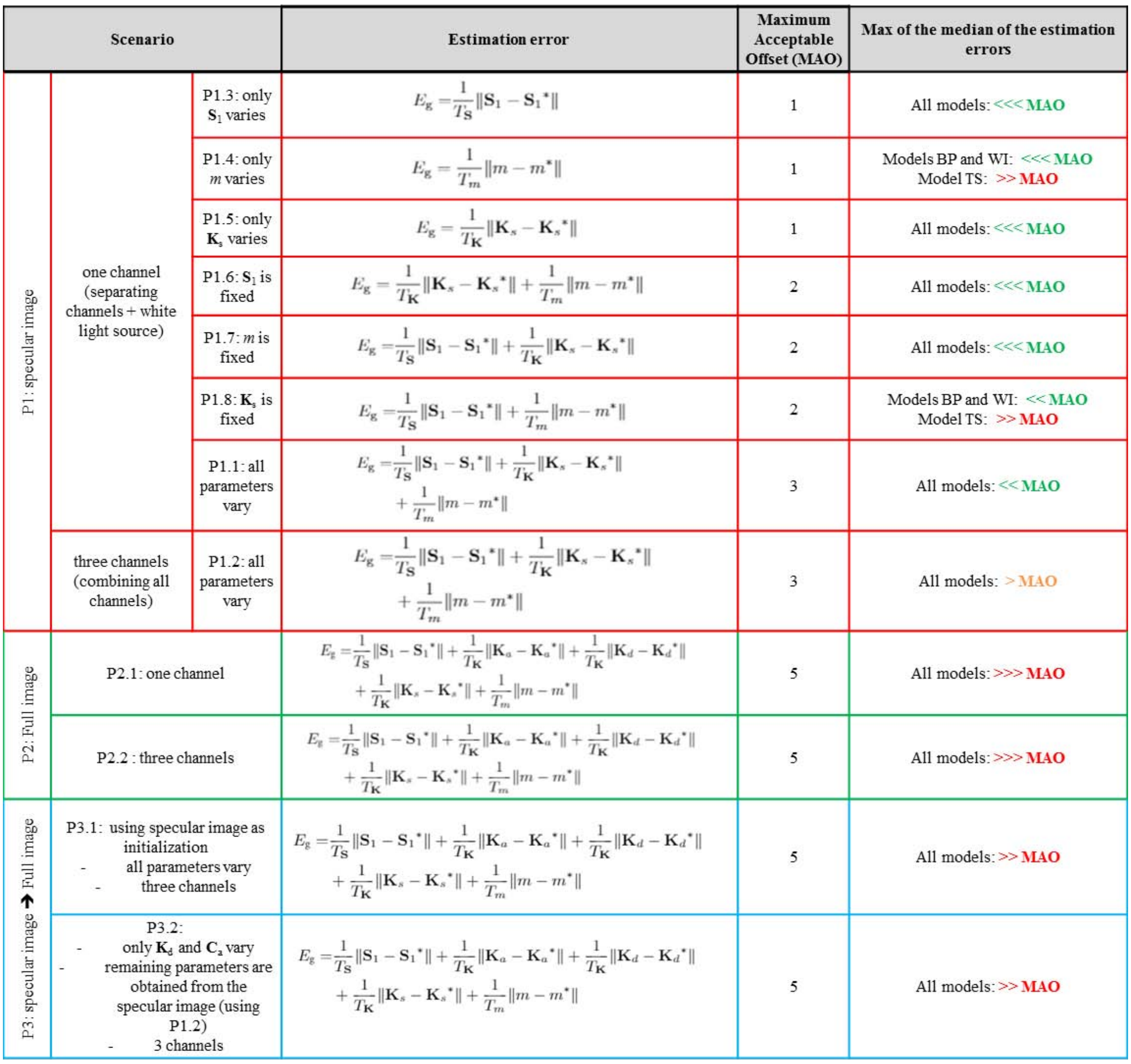
3.4.3 Optimization

Combining All Channels

For the scenarios **P1.2**, **P2.2**, **P3.1** and **P3.2**, the non-linear optimization is performed on the RGB color channels of the image region Ω. We retrieve all parameters by minimizing the following non-linear least squares cost:

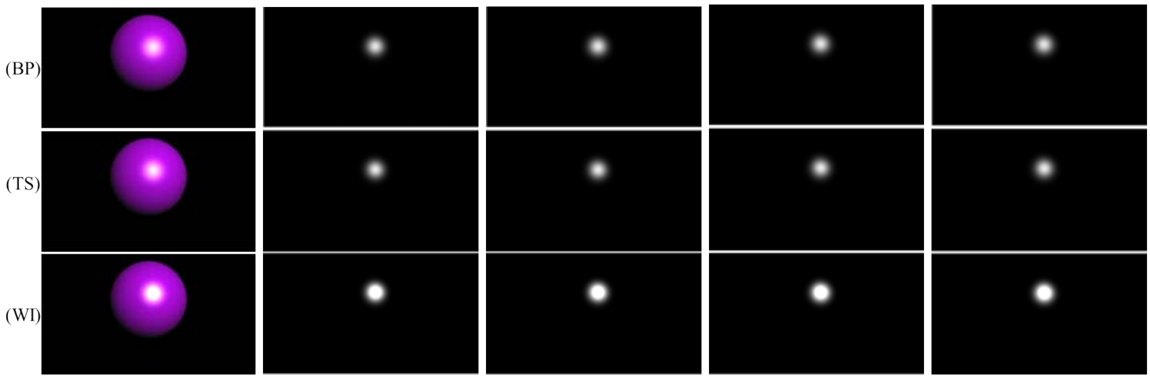
$$\underset{\mathbf{S}_1^*,\mathbf{K}_a^*,\mathbf{K}_d^*,\mathbf{K}_s^*,m^*}{\arg\min} \quad C^2_{\text{photo}} = \frac{1}{|\Omega|}\sum_{\mathbf{P}\in\Omega} \|\mathbf{I}_r(\mathbf{P}) - \mathbf{I}_e(\mathbf{P},\mathbf{S}_1^*,\mathbf{K}_a^*,\mathbf{K}_d^*,\mathbf{K}_s^*,m^*)\|_2^2. \tag{3.4}$$

Letters with an asterisk represent the estimated parameters.



Scenario			Estimation error	Maximum Acceptable Offset (MAO)	Max of the median of the estimation errors
P1: specular image	one channel (separating channels + white light source)	P1.3: only $\mathbf{S}_1$ varies	$E_g = \frac{1}{T_\mathbf{S}}\|\mathbf{S}_1 - \mathbf{S}_1^*\|$	1	All models: <<< MAO
		P1.4: only m varies	$E_g = \frac{1}{T_m}\|m - m^*\|$	1	Models BP and WI: <<< MAO Model TS: >> MAO
		P1.5: only $\mathbf{K}_s$ varies	$E_g = \frac{1}{T_\mathbf{K}}\|\mathbf{K}_s - \mathbf{K}_s^*\|$	1	All models: <<< MAO
		P1.6: $\mathbf{S}_1$ is fixed	$E_g = \frac{1}{T_\mathbf{K}}\|\mathbf{K}_s - \mathbf{K}_s^*\| + \frac{1}{T_m}\|m - m^*\|$	2	All models: <<< MAO
		P1.7: m is fixed	$E_g = \frac{1}{T_\mathbf{S}}\|\mathbf{S}_1 - \mathbf{S}_1^*\| + \frac{1}{T_\mathbf{K}}\|\mathbf{K}_s - \mathbf{K}_s^*\|$	2	All models: <<< MAO
		P1.8: $\mathbf{K}_s$ is fixed	$E_g = \frac{1}{T_\mathbf{S}}\|\mathbf{S}_1 - \mathbf{S}_1^*\| + \frac{1}{T_m}\|m - m^*\|$	2	Models BP and WI: << MAO Model TS: >> MAO
		P1.1: all parameters vary	$E_g = \frac{1}{T_\mathbf{S}}\|\mathbf{S}_1 - \mathbf{S}_1^*\| + \frac{1}{T_\mathbf{K}}\|\mathbf{K}_s - \mathbf{K}_s^*\| + \frac{1}{T_m}\|m - m^*\|$	3	All models: << MAO
	three channels (combining all channels)	P1.2: all parameters vary	$E_g = \frac{1}{T_\mathbf{S}}\|\mathbf{S}_1 - \mathbf{S}_1^*\| + \frac{1}{T_\mathbf{K}}\|\mathbf{K}_s - \mathbf{K}_s^*\| + \frac{1}{T_m}\|m - m^*\|$	3	All models: > MAO
P2: Full image	P2.1: one channel		$E_g = \frac{1}{T_\mathbf{S}}\|\mathbf{S}_1 - \mathbf{S}_1^*\| + \frac{1}{T_\mathbf{K}}\|\mathbf{K}_a - \mathbf{K}_a^*\| + \frac{1}{T_\mathbf{K}}\|\mathbf{K}_d - \mathbf{K}_d^*\| + \frac{1}{T_\mathbf{K}}\|\mathbf{K}_s - \mathbf{K}_s^*\| + \frac{1}{T_m}\|m - m^*\|$	5	All models: >>> MAO
	P2.2 : three channels		$E_g = \frac{1}{T_\mathbf{S}}\|\mathbf{S}_1 - \mathbf{S}_1^*\| + \frac{1}{T_\mathbf{K}}\|\mathbf{K}_a - \mathbf{K}_a^*\| + \frac{1}{T_\mathbf{K}}\|\mathbf{K}_d - \mathbf{K}_d^*\| + \frac{1}{T_\mathbf{K}}\|\mathbf{K}_s - \mathbf{K}_s^*\| + \frac{1}{T_m}\|m - m^*\|$	5	All models: >>> MAO
P3: specular image → Full image	P3.1: using specular image as initialization - all parameters vary - three channels		$E_g = \frac{1}{T_\mathbf{S}}\|\mathbf{S}_1 - \mathbf{S}_1^*\| + \frac{1}{T_\mathbf{K}}\|\mathbf{K}_a - \mathbf{K}_a^*\| + \frac{1}{T_\mathbf{K}}\|\mathbf{K}_d - \mathbf{K}_d^*\| + \frac{1}{T_\mathbf{K}}\|\mathbf{K}_s - \mathbf{K}_s^*\| + \frac{1}{T_m}\|m - m^*\|$	5	All models: >> MAO
	P3.2: - only $\mathbf{K}_d$ and $\mathbf{C}_a$ vary - remaining parameters are obtained from the specular image (using P1.2) - 3 channels		$E_g = \frac{1}{T_\mathbf{S}}\|\mathbf{S}_1 - \mathbf{S}_1^*\| + \frac{1}{T_\mathbf{K}}\|\mathbf{K}_a - \mathbf{K}_a^*\| + \frac{1}{T_\mathbf{K}}\|\mathbf{K}_d - \mathbf{K}_d^*\| + \frac{1}{T_\mathbf{K}}\|\mathbf{K}_s - \mathbf{K}_s^*\| + \frac{1}{T_m}\|m - m^*\|$	5	All models: >> MAO

Table 3.1 Results for the different scenarios. The Maximum Acceptance Offset (MAO), defined in section 5.3, is fixed to the number of terms in the estimation error. Both the median and the mean estimation errors are reported. The maximum median error is compared to the corresponding MAO in the last column.



(a) Original image (b) Distance=1× (c) Distance=2× (d) Distance=10× (e) Distance=100×

Fig. 3.4 An example where the specular highlights do not considerably change when we vary the distance between the light source and the object while maintaining the same light direction.

Separating Channels

In the other scenarios, we perform a non-linear optimization distinctly for each RGB channel. This means that the values of the parameters $\mathbf{K}_s$, $\mathbf{K}_d$ and $\mathbf{C}_a$, are estimated separately per channel. In this case, we minimize the following cost:

$$\underset{\mathbf{S}_1^*,\mathbf{K}_a^*,\mathbf{K}_d^*,\mathbf{K}_s^*,m^*}{\arg\min} \; C^2_{\text{photo}} = \frac{1}{|\Omega|}\sum_{\mathbf{P}\in\Omega}\left(I^c_r(\mathbf{P}) - I^c_e(\mathbf{P},\mathbf{S}_1^*,K_a^{c*},K_d^{c*},K_s^{c*},m^*)\right)^2. \tag{3.5}$$

Each parameter with the notation ^{c} corresponds to its value in the color channel c. In total, three separate minimizations are carried out for each test. We obtain three different values for $\mathbf{S}_1$ and m and use the median as final estimate.

Details

We ran our tests with several optimization algorithms that were used in solving this problem by previous approaches [13, 44, 92, 107]: the downhill simplex, gradient descent and Levenberg-Marquardt. We report the best obtained results among these algorithms which were for Levenberg-Marquardt. We set the initialization values of the parameters by applying a random perturbation on the ground-truth values with different magnitudes, see section 5.5. We set the termination tolerance for the cost value to 10^{-15} and the termination tolerance for the step size to 10^{-20}. The algorithm uses finite differences. It stops without converging if the number of function evaluations is larger than 2000.

3.5 Experimental Protocol

3.5.1 Image Set

We consider a set of synthetic images generated by the corresponding reflectance models. In total, we perform our experiments on 81 images. We consider 3 objects shown in figure 3.2. Each object is characterized by a different type of curvature and roughness in order to exhaustively represent a maximum number of object types. For each object, 9 camera poses are used to diversify the specular highlight configuration. Thus, we use 27 images per object. A sample of these images is shown in figure 3.3. We also test the approach on real images, see figure 3.12. We arrange a real scene with multiple specular objects as shown in figure 3.11. Different types of object are included in the scene. 4 images are taken from different camera poses.

3.5.2 Residual Error

Once Levenberg-Marquardt stops by reaching one of the termination criteria of section 4.3.3, we compute the residual error using the final estimated parameters. This residual is photometric. It assesses the ability of the method to fit the observed image but does not tell us if the estimated parameters match the ground-truth values. To do so, we compute the estimation error.

3.5.3 Estimation Error

The estimation error is fixed as the normalized difference between the true and estimated values of the parameters:

$$\begin{aligned}E_{\text{g}} =&\frac{1}{T_{\mathbf{S}}}||\mathbf{S}_1 - \mathbf{S}_1{}^*||_2 + \frac{1}{T_{\mathbf{K}}}||\mathbf{K}_a - \mathbf{K}_a{}^*||_2 + \frac{1}{T_{\mathbf{K}}}||\mathbf{K}_d - \mathbf{K}_d{}^*||_2\\ &+ \frac{1}{T_{\mathbf{K}}}||\mathbf{K}_s - \mathbf{K}_s{}^*||_2 + \frac{1}{T_m}||m - m^*||_2,\end{aligned} \tag{3.6}$$

where $T_{\mathbf{S}}$, $T_{\mathbf{K}}$ and T_m are weights computed independently from the numerical error of each parameter in equation (3.6). Since the parameters have different orders of magnitude, this allows us to normalize the terms to a common scale. We define MAO, the Maximum Acceptance Offset as the maximum tolerated estimation error. Each term of equation (3.6) has to be lower than 1. So, MAO is 5 for **P2** and **P3** (5 terms) and it is 3, 2 or 1 for **P1** as $\mathbf{K}_a$ and $\mathbf{K}_d$ are not estimated (see table 3.1). The estimation error and MAO allow us to evaluate the performance of each model and to inform us if the optimization algorithm converged to the true solution. For the different scenarios on synthetic images, we give the specific estimation error in table 3.1. For real images, only the ground-truth of the light position is known and we thus only use the first term of equation (3.6).

3.5.4 Weights Computation

The weights $T_{\mathbf{S}}$, $T_{\mathbf{K}}$ and T_m depend on the orders of magnitude of the estimated parameters and are used to normalize the estimation error (3.6). These weights also provide information on the size of the convergence basin and error tolerance. To determine their values, we launch the optimization algorithm on the images by initializing the parameters with the ground-truth values. The weight T_S is then determined by:

$$T_{\mathbf{S}} = \mu_{\mathbf{S}} + \alpha_{\mathbf{S}}\sigma_{\mathbf{S}}, \tag{3.7}$$

where $\mu_{\mathbf{S}}^2 = \frac{1}{j}\sum_{i=1}^{j} \|\mathbf{S}_1(i) - \mathbf{S}_1{}^*(i)\|_2^2$ is the mean square error of the light source's position on $j = 300$ trials, $\sigma_{\mathbf{S}}$ the standard deviation and $\alpha_{\mathbf{S}}$ a coefficient chosen so that 95% of the trials satisfy the following condition:

$$\mu_{\mathbf{S}} - \alpha_{\mathbf{S}}\sigma_{\mathbf{S}} \leq \|\mathbf{S}_1(i) - \mathbf{S}_1(i)^*\| \leq \mu_{\mathbf{S}} + \alpha_{\mathbf{S}}\sigma_{\mathbf{S}}. \tag{3.8}$$

This condition allows us to exclude high values of the error. The same method is used for T_K and T_m. The values retrieved in this test depend on the orders of magnitude of the estimated parameters.

Magnitude index	perturbation on $\mathbf{S}_1$	perturbation on $\mathbf{K}_a, \mathbf{K}_d, \mathbf{K}_s$	perturbation on m in BP	perturbation on m in TS and WI
0	0.0	0.0	0.0	0.0
1	1.0	0.010	1.0	0.005
2	1.66	0.016	1.66	0.008
3	2.78	0.027	2.78	0.013
4	4.64	0.046	4.64	0.023
5	7.74	0.077	7.74	0.038
6	12.91	0.129	12.91	0.064
7	21.54	0.215	21.54	0.107
8	35.93	0.359	35.93	0.179
9	59.94	0.599	59.94	0.299
10	100.0	1.0	100.0	0.500

Table 3.2 Perturbation magnitudes on the ground-truth for the initialization of each parameter.

3.5.5 Initialization

In our tests, the initialization of the light position, the reflectance parameters and the roughness parameter is provided by a perturbation of the ground-truth parameters with a varying magnitude. We set 10 perturbation magnitudes in a logarithmic scale for each parameter. We use the Euclidean distance between the initialization value and the true value of the parameters. For example, if the perturbation magnitude of the light position is $t = 10$ then the initialization value $\mathbf{S}_1{}^*$ satisfies the condition $||\mathbf{S}_1 - \mathbf{S}_1{}^*||_2 = 10$. Table 3.2 lists the perturbation magnitudes that we set for the different parameters. For each magnitude, we generate 30 sets of random values. Therefore, a total of 300 different tests are performed on each image in the dataset. For each parameter, the generated initial values guarantee a fixed estimation error that corresponds to the value in table 3.2. For real images, only the light position is known in our experiments. So, the remaining ground-truth values of the model parameters are set manually. We empirically approximate the best values of these parameters that visually recreate the reference images using the corresponding reflectance models.

3.6 Results

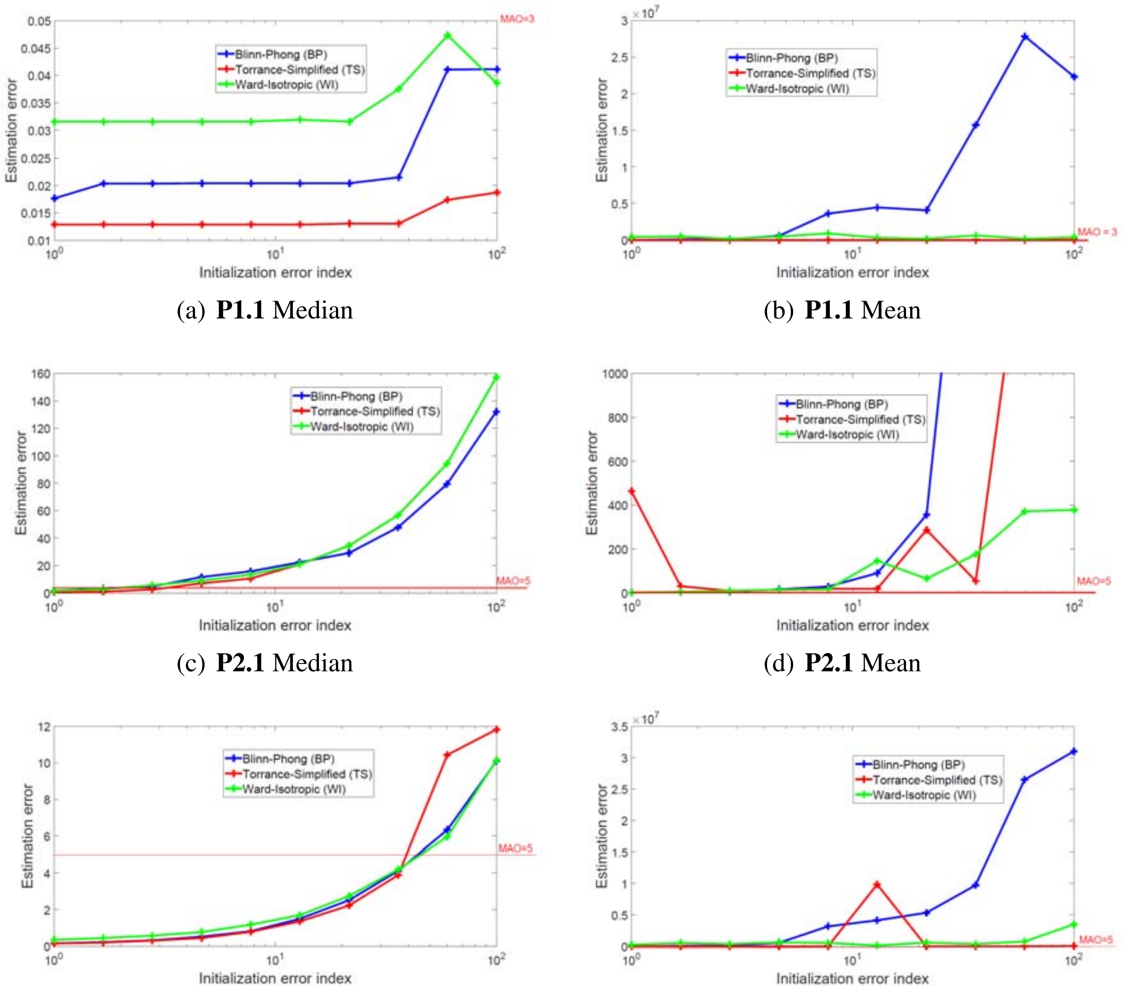
In this section, we report the inverse local reflectance results on both synthetic and real data. A predicted specular component image is synthesized using the estimated parameters and compared to ground truth images.

3.6.1 Weights

In table 3.3, we present the weights defined in section 5.4 for each reflectance model. They were calculated as explained in section 5.4.

Weights	BP	TS	WI
$T_{\mathbf{S}}$	0.207	3.620	0.600
$T_{\mathbf{K}}$	0.154	0.160	0.180
T_m	0.557	0.004	0.004

Table 3.3 The weight per reflectance model.



(a) **P1.1** Median

(b) **P1.1** Mean

(c) **P2.1** Median

(d) **P2.1** Mean

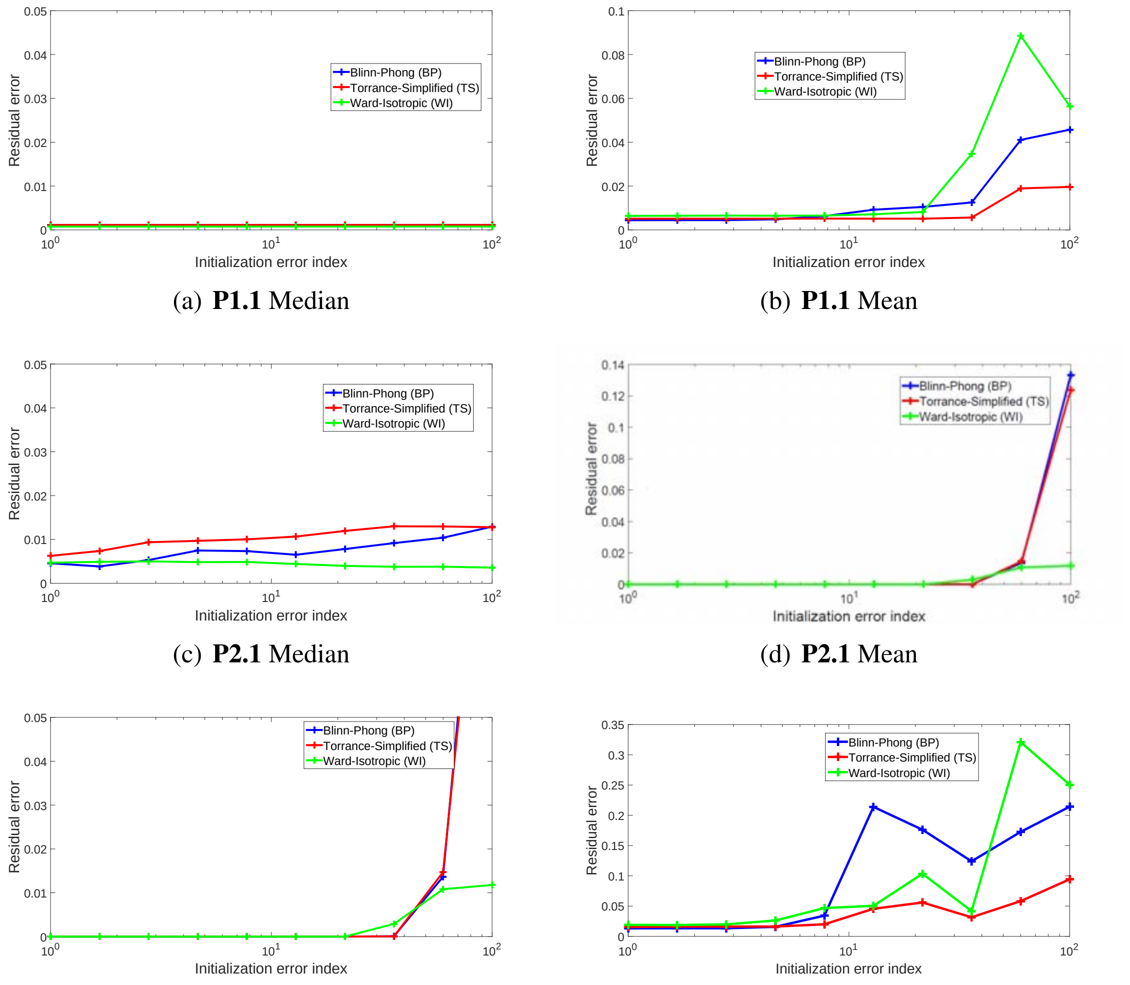
(e) **P3.1** Median

(f) **P3.1** Mean

Fig. 3.5 *Estimation errors* of the scenarios **P1.1**, **P2.1** and **P3.1**. We report both median and mean of the errors for each initialization magnitude.

3.6.2 Synthetic Data

For each image of the dataset, we launch a total of 300 trials (10 initialization magnitudes × 30 random values). As shown in table 3.2, the initialization values are increasingly distant from the ground-truth. The estimation errors are then retrieved as shown in figure 3.5. We compare the three models with the same scale using the mean and the median on the trials per perturbation magnitude. According to the results in table 3.1, we see that the first scenario **P1** is a well-posed problem for models BP and WI and under different conditions. The model TS gives slightly different results for scenarios **P1.4** and **P1.8**. So, with a single specular highlight, we can robustly estimate the parameters of the specular component provided that we have separated the diffuse and the specular components of the original image. We



(a) **P1.1** Median

(b) **P1.1** Mean

(c) **P2.1** Median

(d) **P2.1** Mean

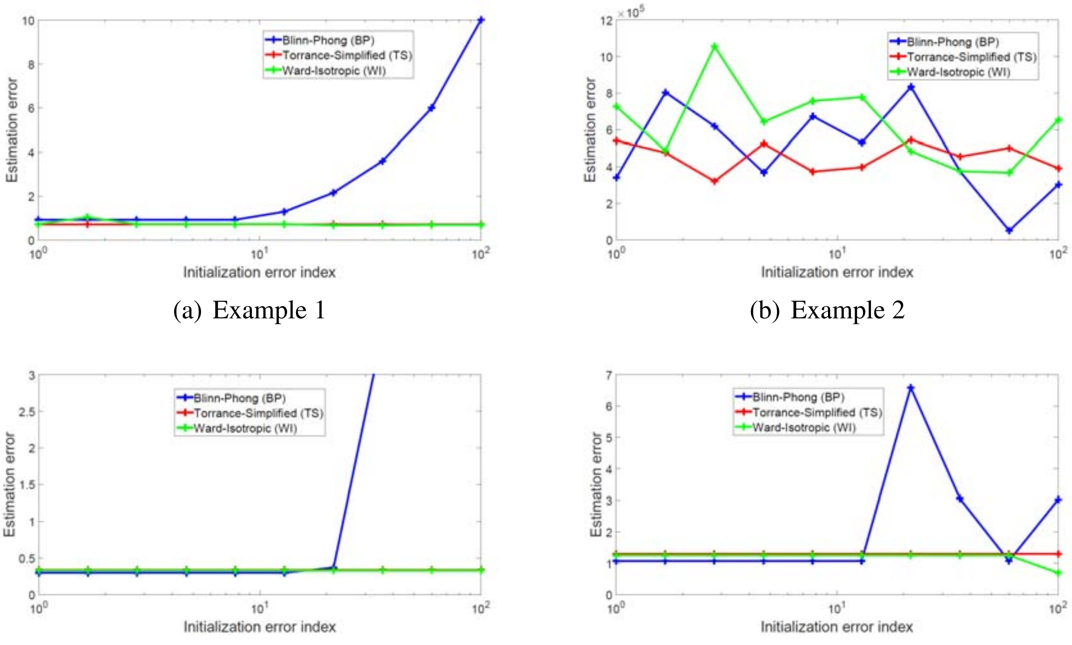
(e) **P3.1** Median

(f) **P3.1** Mean

Fig. 3.6 *Residual errors* of the scenarios **P1.1**, **P2.1** and **P3.1** on synthetic data. We report the median of the errors for each initialization magnitude.

also note that the results deteriorate when we combine the three channels. This means that separating channels is more robust than combining them.

Once we consider the cases **P2** of retrieving the parameters from the original image, the problem becomes ill-posed since we obtain high estimation errors with respect to MAO. When combining the two types of input data in **P3**, we observe that the problem remains ill-posed. In fact, when we consider the specular component results in **P3.1** and **P3.2**, we see that the estimation of $\mathbf{K}_s$, $\mathbf{C}_a$ is still difficult. This is mainly because we consider a specular highlight region where the diffuse component is outweighted by the specular component. A solution would be to estimate these two parameters separately from a diffuse region of the image.



(a) Example 1

(b) Example 2

(c) Example 3

(d) Example 4

Fig. 3.7 The real data estimation errors on the four examples. The input images are shown in figure 3.12.

According to the mean estimation error reported in figure 3.5, some examples are harder to estimate than others. We show one of the hardest examples in figure 3.4. We see that varying the light distance while fixing the remaining parameters generates approximately the same image. This explains why the algorithm fails to estimate the real position of the light source. The residual errors shown in figure 3.6 tell us that the algorithm converges to a solution that fits the considered models for the all scenarios. This solution is however different than ground truth in scenarios **P2.1** and **P3.1**.

3.6.3 Real Data

Starting from our analysis of the results on the synthetic data, we apply scenario **P1.1** to validate our findings on real data. To separate the diffuse and specular components, we use two polarizers, one in front of the camera and another in front of the light source. We obtain two images, a first with the polarizers' angles being parallel and a second with their angles being orthogonal. This generates the original image and the diffuse image (without the specular component). The specular component image is obtained by subtracting the two images. In figure 3.12, we show the real data used to test our approach and their corresponding specular component. The 3D surface of the scene is reconstructed using the

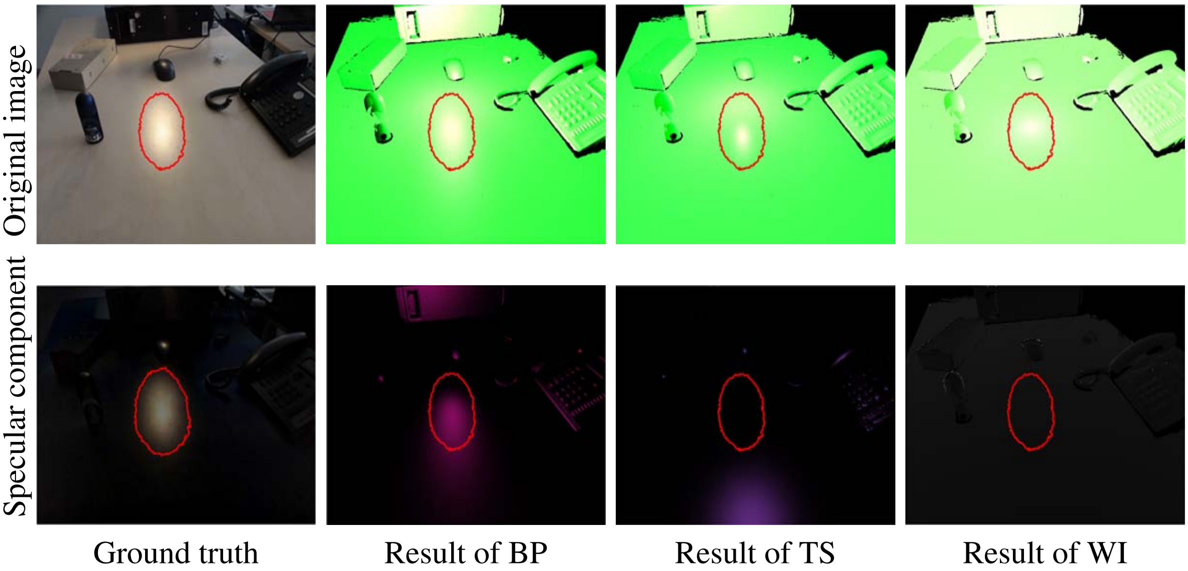


Fig. 3.8 Results on the real dataset using scenario **P2.1**. We generate the three components in the first row using all the estimated parameters. We generate the specular component in the second row using only parameters $\mathbf{S}_1$, $\mathbf{K}_s$ and m. The optimization algorithm reaches a local minimum after exceeding the maximum number of iterations which results in high residual error. This explains the erroneous color (green) in the top images. It corresponds to the estimated diffuse coefficient $\mathbf{K}_d$.

HandySCAN 3D scanner from Creaform[2] as shown in figure 3.11. We manually define 3D-2D correspondences between surface points and pixels in each image. Then, we perform an iterative minimization of the re-projection error using Levenberg-Marquardt in order to estimate the camera pose. The camera intrinsic parameters are obtained using pre-calibration. The distortion in the images is corrected and the resolution is 640 × 480 pixels.

Since the ground-truth values of the light intensity, reflectance parameters and roughness are unknown, we manually generate synthetic images with different parameters until we obtain a result that is close enough to the reference images. These values are used only to set the initialization values but not to compute the estimation error. The median of the estimation errors of scenario **P1.1** for each real example are reported in figure 3.7. We follow the same experimental scheme as for the synthetic data: 10 initialization magnitudes and 30 trials per magnitude. The results of examples 1, 3 and 4 show that the algorithm converges to a stable solution. TS and WI models give a stable solution which is maintained even for high initialization magnitudes. However, BP model maintains a stable solution only for initialization magnitudes less than 10. This solution is very close to the true light position. The small error is mainly due to noise from the camera pose, the 3D reconstruction and the

[2] www.creaform3d.com/en

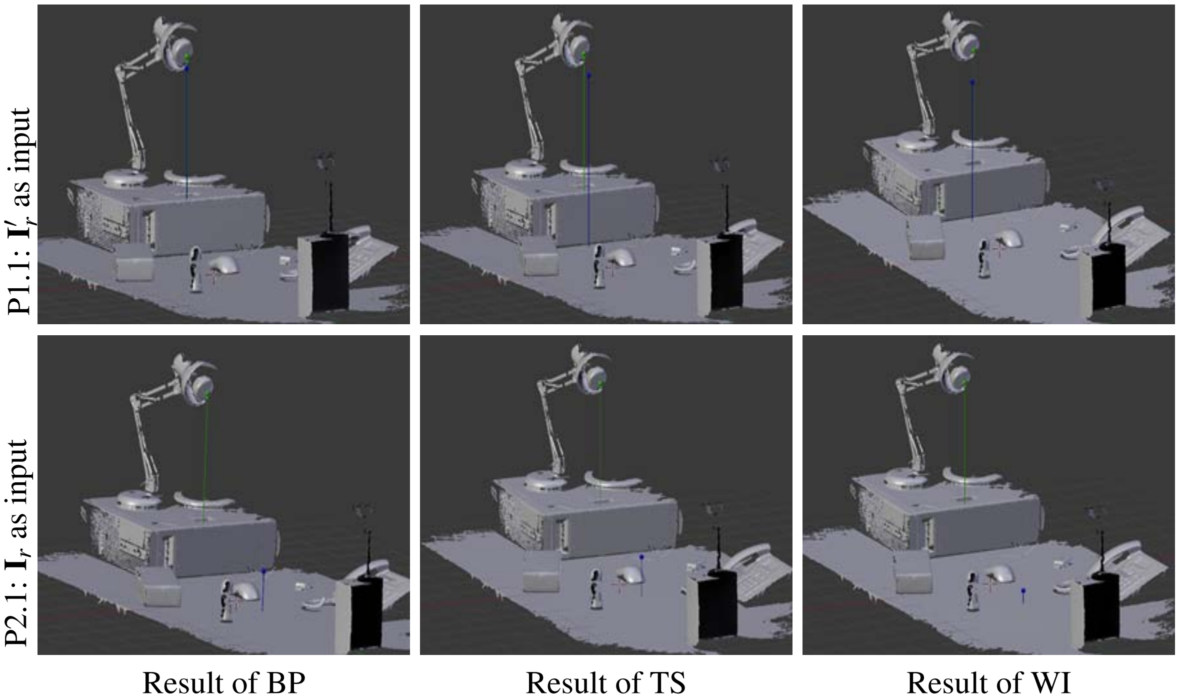
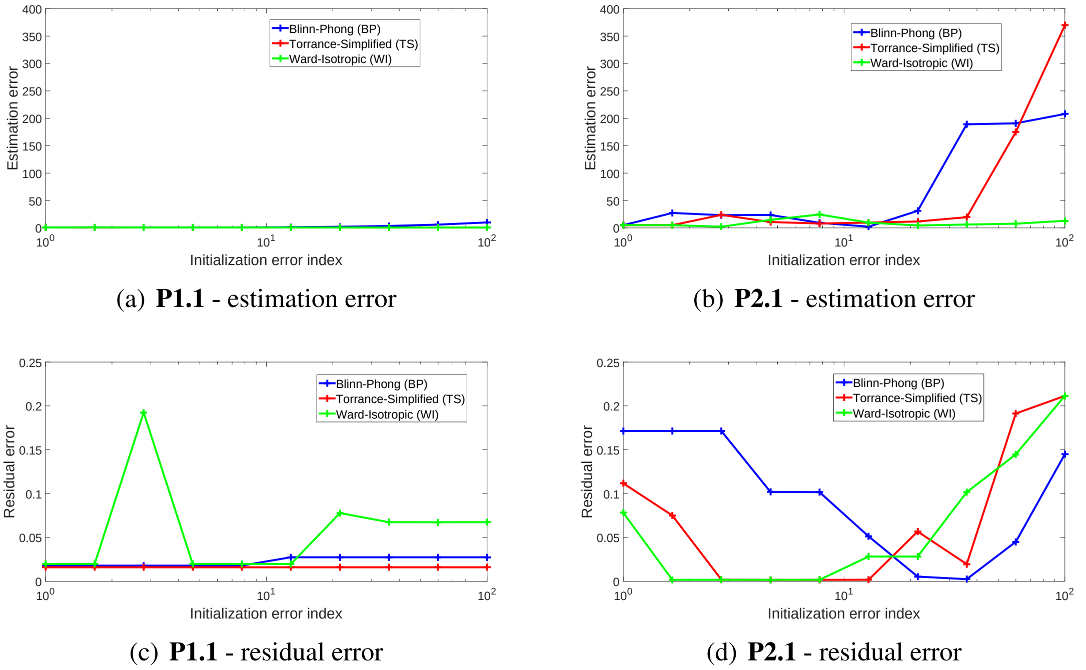


Fig. 3.9 Comparing the light position estimation results of scenarios **P1.1** and **P2.1** on example 1. The true position is indicated by the green sphere and the estimated position is indicated by the blue sphere

diffuse specular separation. Example 2 gives high estimation errors for the three models. However as seen in figure 3.12, the models BP and TS are well-fitted for example 2. This is actually in concordance with the special case observed in the synthetic data from figure 3.4. The re-synthesized specular components shown in figure 3.12 are very close to the ground truth for the models BP and TS for all examples. For the model WI, even though the light position is well-estimated (see figure 3.7), it seems that the remaining parameters $\mathbf{K}_s$ and m are false. We also tested scenario **P2.1** on real data. The results of example 1 are reported in figure 3.8 and 3.10. We compare in figure 3.9 the estimated light position on example 1 using scenarios **P1.1** and **P2.1** with the same initialization values. In figure 3.10, the estimation error is very high and the retrieved solution is unstable depending on the initialization magnitude. This agrees with the results on synthetic data. By generating images using the estimated parameters, we verify that the algorithm fails to fit the real data using the model with the three components in contrast to the model with only the specular component. To sum up, these results confirm our findings on synthetic data: we can robustly estimate the reflectance model's specular parameters using a single-spot approach provided that a diffuse-specular separation is achieved. The model TS gives the smallest estimation errors on most examples and is stabler than other models. Some examples make the local



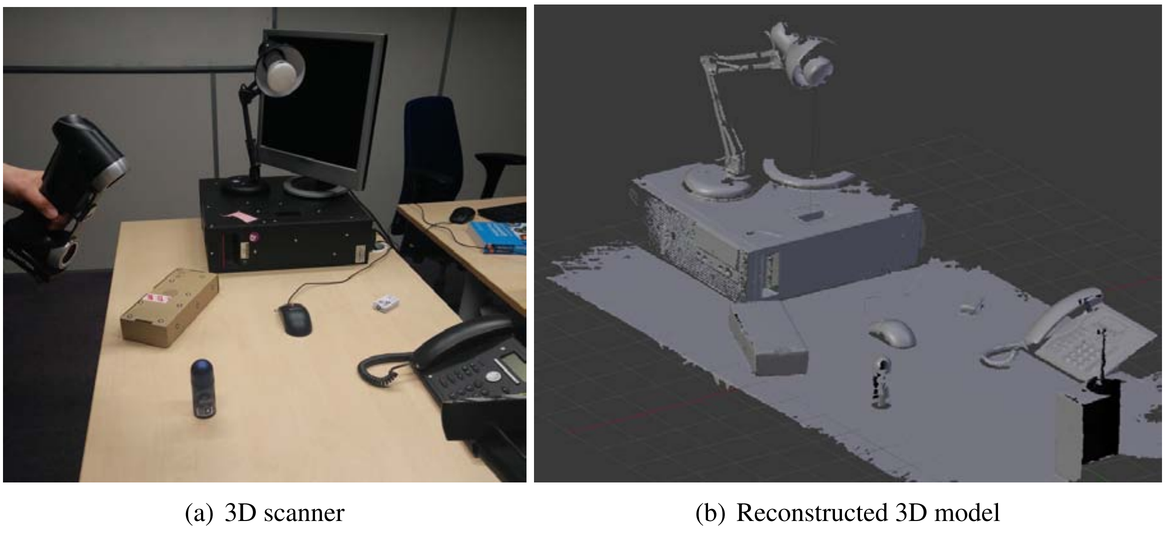
(a) **P1.1** - estimation error

(b) **P2.1** - estimation error

(c) **P1.1** - residual error

(d) **P2.1** - residual error

Fig. 3.10 Comparing the estimation and residual errors of scenarios **P1.1** and **P2.1** on real data. We show the median of the errors on example 1 depending on the initialization magnitude.



(a) 3D scanner

(b) Reconstructed 3D model

Fig. 3.11 Illustrating the experimental setup for the acquisition of real data ground truth.

reflectance inversion hard for this approach. These include the case where the shape of the specularity is weakly influenced by the distance to the light source as in figure 3.4.

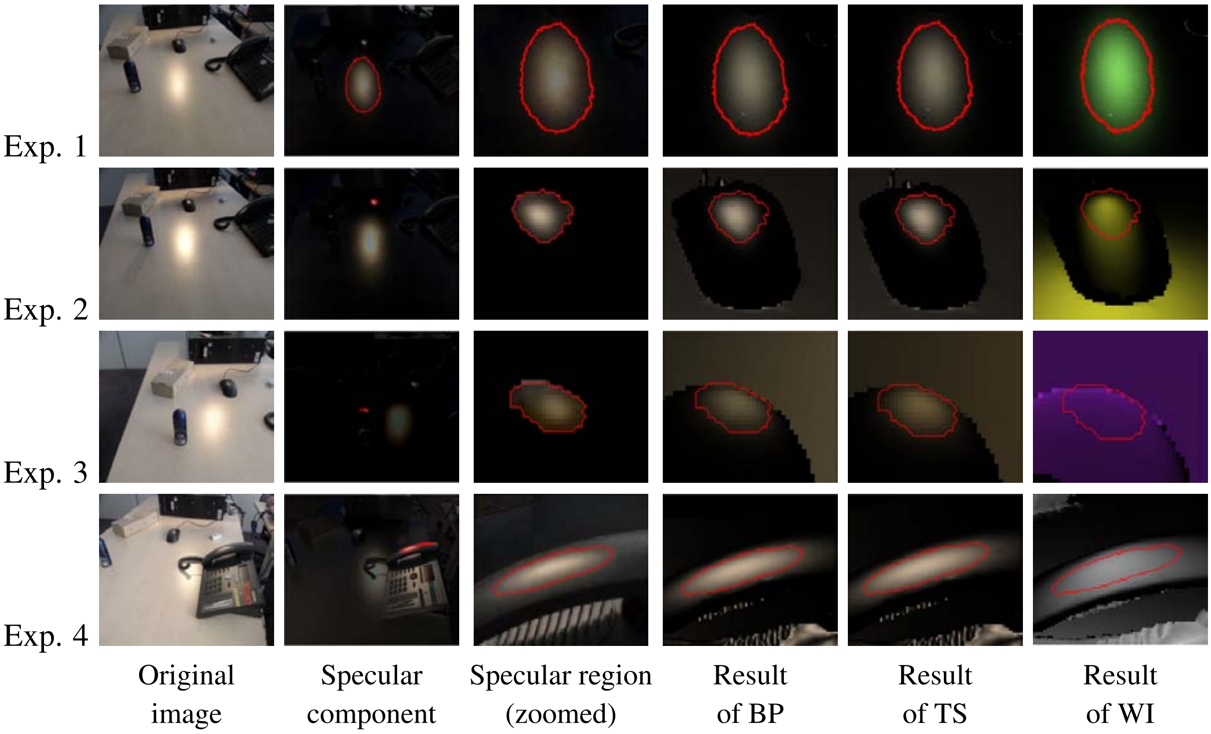


Fig. 3.12 Results on the real dataset using scenario **P1.1**. The specular image obtained after subtracting the diffuse image from the original image is in the second column. A zoom on the considered highlight region is in the third column. The generated images using the estimated parameters from BP, TS and WI are, respectively, in the fourth, the fifth and the sixth columns. The faulty colors in the results of WI are due to the flawed estimation of the reflectance coefficient $\mathbf{k}_s$ which is very sensitive to the choice of the color space.

3.7 Discussion

The results in this chapter are obtained on test scenarios using a single point light source, known surface geometry and homogeneous isotropic surfaces. Although these assumptions may seem simplistic, they allow us to focus on the main objective of our study which is the correlation between the input data and the efficiency of the existing light and material reconstruction methods. They also form a basis scenario for later studies on more complex light configurations or surfaces. Importantly, this work is also the first to quantitatively investigate the relevance of specular highlights for solving the inverse local reflectance problem.

3.8 Conclusion

We have addressed the problem of local reflectance inversion from a single specular highlight which we call the single-spot approach. We exhaustively evaluated this approach with different scenarios. In the light of our findings, we recommend the single-spot approach as a flexible method to local reflectance recovery in AR and DR. We showed that a specular-diffuse separation is an essential step to ensure the solvability of this problem. This approach can be used in applications such as AR and DR without the need of any priors on the number of light sources in the scene since each specularity is processed separately. The fact that we can find the position of a 3D point in space, the light position, from a single observation, opens many perspectives for similar inversion problems like camera localization. The specular highlight properties will be further investigated in this direction. In future work, we will carry out an analysis of the robustness of local inversion in the presence of significantly noisy data. We will also investigate the combination with methods for separating the diffuse and specular components using the latest deep learning approaches.

Chapter 4

Conclusion and Future Work

4.1 Conclusion

This thesis addressed the technique of Diminished Reality. In particular, we focused on the update of the pixels' intensities in the target region when the camera moves. Knowing that this problem is caused by the specular component in the image, we addressed it differently through two approaches. While our first approach falls into the category of specularity modeling, our second approach falls into the category of illumination model inversion.

4.1.1 Specularity Propagation

Our first approach is based on the hypothesis that the pixel intensities are altered solely by the specularity crossing in the target region. This hypothesis relies on the local illumination models commonly used in Computer Graphics. They clearly consider that any intensity change in the image of a static scene between two view-points is caused by the specular component of the image. Our approach is a specularity modeling method applied for DR. Our method used the TPS to interpolate missing pixels intensities in the target region. By applying the TPS, we were able to propagate the change in intensities from the outside to the inside of the target region ensuring the smoothness property of specularities. We established that this property is usually true for convex and planar surfaces in the presence of point light sources. We also proposed an extension of our approach that takes into consideration further properties of the specularity for planar surfaces. For instance, this extension constrains the intensity isocontours of the specularity to ellipses. We showed that the elliptical property is valid according to the local illumination models as well. As a result, our real-time solution outperforms state-of-the-art results in terms of rendering quality on several real examples. This work was published in [40, 41]. This first approach has the limitation that a part of the specularity has to be visible outside the target region. To overcome this, we addressed it as a problem of illumination model inversion.

4.1.2 Illumination Model Inversion

Our second approach explored the concept of illumination model inversion assuming that a local illumination model is valid for the observed scene. The exhaustive evaluation proposed in chapter 3 offers an overview of the minimum requirements to achieve a local illumination inversion. In the light of our findings, a single specular region can be exploited to recover the scene's specular parameters and the position of the light source. In the presence of multiple light sources, analyzing each specular region separately allows one to avoid the ambiguity behind associating each specularity to its corresponding light source. This facilitates the problem of illumination model inversion in the presence of more than a single light source. This work was published in [42].

4.2 Future Work

4.2.1 Diminished Reality in the Presence of Multiple Materials

In the presence of multiple materials, the specular highlights are altered depending on the roughness and specular parameters of each surface. In fact, the specularity isocontours will be shifting at the material transition boundaries of each material. This would be a limitation for our extended specularity propagation model with elliptical constraints (CPM). Therefore, it is necessary to segment the target region and estimate the specularity propagation depending on the properties of the material. Several segmentation methods could be considered for this purpose, especially with recent advances in deep learning methods [21]. During this thesis, we also established the efficiency of spectral segmentation of materials in the task of image completion [15, 14]. Similarly, multi-spectral image segmentation could be also exploited for specularity propagation. Further details on this work are given in the appendix.

4.2.2 Highlight Ovals

The recent work in [5] showed that according to Phong's local illumination model, the specular highlights isocontours can be better modeled by Highlight Ovals, a model of plane ovals based on two foci defined in the 3D space. These ovals approximate better the specularity's shape for grazing angles between the surface and the camera as shown in figure 4.1. So, our model can be further improved by including this particular model of highlight

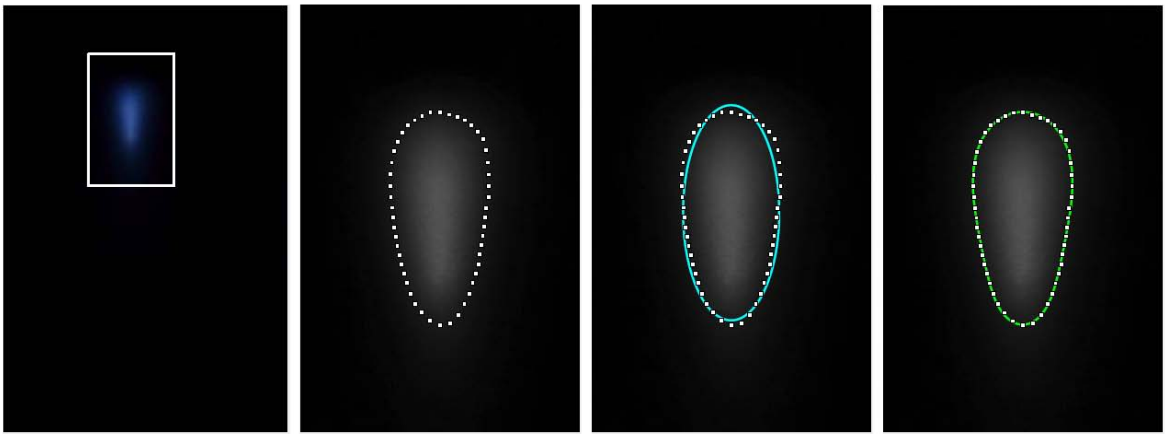


Fig. 4.1 Comparison between ellipse model fitting and highlight ovals model fitting of a real image of a specularity. (from left to right) The image was taken by a digital camera in fast-shutter mode, allowing the sensor to image the specular highlight. The box around the specular highlight has size 404x616 pixels. The sample points were extracted from the image. The ellipse and algebraic highlight oval obtained fitting residuals of 3.26 px and 0.72 px respectively. Images taken from [5].

ovals. In future work, by investigating other illumination models, e.g. Blinn-Phong, and convex surfaces we could approximate the isocontours of specular highlight by other shapes in the same spirit as the Highlight Ovals.

4.2.3 Specular-Diffuse Separation

In chapter 2, we showed that the separation of the diffuse and specular components is a crucial step in the task of illumination model inversion. Currently, single-image methods mostly rely on the Dichromatic Reflection Model and the fact that specularities retain the illumination's color to do the separation. However, the separation problem with a single image being ill-posed because of the ambiguity of the image formation process [63], they make strong assumptions about the scene such as a single illumination of known color, no saturated pixels and linear response of the capture device. This obviously hinders the generic applicability of the methods. Therefore, this is still a challenging and open problem. In future work, we could investigate a deep learning approach to overcome the limitations in applicability. The idea is that the network will work out the intricate relationships between an image and its diffuse part. Recently, promising work was proposed in this direction [72, 30, 79, 95].

4.2.4 Initialization in Local Illumination Model Inversion

As mentioned in chapter 3, our method uses non linear optimisation to estimate the parameters of different local illumination models. Currently, we initialize the optimisation algorithm with random values. In the future, we can set the initial values using the analytic approach in [5] that introduces the highlight ovals model for specularities. Therefore, we could drastically speed up the estimation of these parameters. We have already proven that good initialization allows better estimation results. Combining this analytic approach as initialization with the illumination model inversion discussed in chapter 2 will allow us to propose a complete solution for specularity prediction. A first step would be estimating the model's parameters and light position from a specularity on a planar surface. Then, this will allow us to predict the specularity's shape on more complex surfaces of the same material, assuming that the scene's geometry is known.

References

[1] Aittala, M. (2010). Inverse lighting and photorealistic rendering for augmented reality. *The Visual Computer*, 26:669–678.

[2] Ballester, C., Bertalmio, M., Caselles, V., Sapiro, G., and Verdera, J. (2001). Filling-in by joint interpolation of vector fields and gray levels. *IEEE transactions on image processing*, 10(8):1200–1211.

[3] Barnes, C., Shechtman, E., Finkelstein, A., and Goldman, D. (2009). Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (TOG)*, 28(3):24.

[4] Barron, J. T. and Malik, J. (2015). Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1670–1687.

[5] Bartoli, A. (2019). The highlight ovals. *Journal of Mathematical Imaging and Vision*, pages 1–25.

[6] Bell, S., Upchurch, P., Snavely, N., and Bala, K. (2015). Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3479–3487.

[7] Bertalmio, M., Bertozzi, A. L., and Sapiro, G. (2001). Navier-stokes, fluid dynamics, and image and video inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–355. IEEE.

[8] Bertalmio, M., Sapiro, G., Caselles, V., and Ballester, C. (2000). Image inpainting. In *ACM Special Interest Group on GRAPHics and Interactive Techniques (SIGGRAPH)*, pages 417–424. ACM.

[9] Blake, A. and Bülthoff, H. (1990). Does the brain know the physics of specular reflection? *Nature*, 343(6254):165–168.

[10] Blinn, J. F. (1977). Models of light reflection for computer synthesized pictures. In *ACM Special Interest Group on GRAPHics and Interactive Techniques (SIGGRAPH)*, volume 11, pages 192–198. ACM.

[11] Boivin, S. and Gagalowicz, A. (2001). Image-based rendering of diffuse, specular and glossy surfaces from a single image. In *ACM SIGGRAPH*, pages 107–116.

[12] Bookstein, F. L. et al. (1989). Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence (PAMI)*, 11(6):567–585.

[13] Boom, B., Orts-Escolano, S., Ning, X., McDonagh, S., Sandilands, P., and Fisher, R. B. (2013). Point light source estimation based on scenes recorded by a rgb-d camera. In *BMVC*.

[14] Bousefsaf, F., Tamaazousti, M., Said, S. H., and Michel, R. (2017). Complétion d'image exploitant des données multispectrales. *Revue Française de Photogrammétrie et de Télédétection*, 215:65–79.

[15] Bousefsaf, F., Tamaazousti, M., Said, S. H., and Michel, R. (2018). Image completion using multispectral imaging. *IET Image Processing*, 12(7):1164–1174.

[16] Buteau, P.-E. and Saito, H. (2015). [poster] retrieving lights positions using plane segmentation with diffuse illumination reinforced with specular component. In *IEEE ISMAR*, pages 202–203. IEEE.

[17] Cahill, N. D., Chew, S. E., and Wenger, P. S. (2015). Spatial-spectral dimensionality reduction of hyperspectral imagery with partial knowledge of class labels. In *SPIE Defense+ Security*, page 94720S.

[18] Cao, F., Gousseau, Y., Masnou, S., and Pérez, P. (2011). Geometrically guided exemplar-based inpainting. *SIAM Journal on Imaging Sciences*, 4(4):1143–1179.

[19] Chakrabarti, A. and Zickler, T. (2011). Statistics of real-world hyperspectral images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 193–200.

[20] Chan, T. F. and Shen, J. (2001). Nontexture inpainting by curvature-driven diffusions. *Journal of Visual Communication and Image Representation (VCIR)*, 12(4):436–449.

[21] Cimpoi, M., Maji, S., and Vedaldi, A. (2014). Deep convolutional filter banks for texture recognition and segmentation. *arXiv preprint arXiv:1411.6836*.

[22] Cook, R. L. and Torrance, K. E. (1982). A reflectance model for computer graphics. *ACM Transactions on Graphics (TOG)*, 1(1):7–24.

[23] Criminisi, A., Perez, P., and Toyama, K. (2003). Object removal by exemplar-based inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–721. IEEE.

[24] Debevec, P., Gortler, S., McMillan, L., Szeliski, R., and Bregler, C. (1998). Image-based modeling and rendering. *ACM SIGGRAPH*, 15.

[25] Dorsey, J. (1995). Radiosity and global illumination. *The Visual Computer*, 11:397–398.

[26] Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in sobolev spaces. In *Constructive Theory of Functions of Several Variables*, pages 85–100. Springer.

[27] Einabadi, F. and Grau, O. (2015). Discrete light source estimation from light probes for photorealistic rendering. In *BMVC*.

[28] Enomoto, A. and Saito, H. (2007). Diminished reality using multiple handheld cameras. In *Proc. ACCV*, volume 7, pages 130–135. Citeseer.

[29] Fitzgibbon, A., Pilu, M., and Fisher, R. B. (1999). Direct least square fitting of ellipses. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 21(5):476–480.

[30] Funke, I., Bodenstedt, S., Riediger, C., Weitz, J., and Speidel, S. (2018). Generative adversarial networks for specular highlight removal in endoscopic images. In *Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 10576, page 1057604. International Society for Optics and Photonics.

[31] Gardner, M.-A., Sunkavalli, K., Yumer, E., Shen, X., Gambaretto, E., Gagné, C., and Lalonde, J.-F. (2017). Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090*.

[32] Geelen, B., Tack, N., and Lambrechts, A. (2014). A compact snapshot multispectral imager with a monolithically integrated per-pixel filter mosaic. In *SPIE MOEMS-MEMS*, page 89740L.

[33] Gendrin, A., Mangold, N., Bibring, J.-P., Langevin, Y., Gondet, B., Poulet, F., Bonello, G., Quantin, C., Mustard, J., Arvidson, R., et al. (2005). Sulfates in martian layered terrains: the omega/mars express view. *Science*, 307(5715):1587–1591.

[34] Georgoulis, S., Rematas, K., Ritschel, T., Gavves, E., Fritz, M., Van Gool, L., and Tuytelaars, T. (2018). Reflectance and natural illumination from single-material specular objects using deep learning. *IEEE transactions on pattern analysis and machine intelligence*, 40(8):1932–1947.

[35] Goldman, D. B., Curless, B., Hertzmann, A., and Seitz, S. M. (2010). Shape and spatially-varying brdfs from photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1060–1071.

[36] Gordon, H. R. and Morel, A. Y. (2012). *Remote assessment of ocean color for interpretation of satellite visible imagery: A review*, volume 4. Springer Science & Business Media.

[37] Green, A. A., Berman, M., Switzer, P., and Craig, M. D. (1988). A transformation for ordering multispectral data in terms of image quality with implications for noise removal. *IEEE Transactions on Geoscience and Remote Sensing*, 26(1):65–74.

[38] Guillemot, C. and Le Meur, O. (2014). Image inpainting: Overview and recent advances. *IEEE signal processing magazine*, 31(1):127–144.

[39] Guo, X., Huang, X., Zhang, L., Zhang, L., Plaza, A., and Benediktsson, J. A. (2016). Support tensor machines for classification of hyperspectral remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 54(6):3248–3264.

[40] Hadj Said, S., Tamaazousti, M., and Bartoli, A. (2017). Image-based models for specularity propagation in diminished reality. *IEEE Transactions on Visualization and Computer Graphics*.

[41] Hadj-Said, S., Tamaazousti, M., and Bartoli, A. (2017). Un modèle de propagation de spécularité dans une vidéo pour la réalité diminuée.

[42] Hadj-Said, S., Tamaazousti, M., and Bartoli, A. (2019). Can we invert a local reflectance model from a single specular highlight with known scene geometry and camera pose?

[43] Hagen, N. and Kudenov, M. W. (2013). Review of snapshot spectral imaging technologies. *Optical Engineering*, 52(9):090901.

[44] Hara, K., Nishino, K., et al. (2005). Light source position and reflectance estimation from a single view without the distant illumination assumption. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4):493–505.

[45] He, K. and Sun, J. (2012). Statistics of patch offsets for image completion. In *European Conference on Computer Vision (ECCV)*, pages 16–29. Springer.

[46] He, K. and Sun, J. (2014). Image completion approaches using the statistics of similar patches. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36(12):2423–2435.

[47] Herling, J. and Broll, W. (2012). Pixmix: A real-time approach to high-quality diminished reality. In *IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 141–150. IEEE.

[48] Herling, J. and Broll, W. (2014a). High-quality real-time video inpainting with pixmix. *IEEE Transactions on Visualization and Computer Graphics*, 20(6):866–879.

[49] Herling, J. and Broll, W. (2014b). High-quality real-time video inpainting with pixmix. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, page 1.

[50] Huang, J.-B., Kang, S. B., Ahuja, N., and Kopf, J. (2014). Image completion using planar structure guidance. *ACM Transactions on Graphics*, 33(4):129.

[51] Huang, J.-B., Kopf, J., Ahuja, N., and Kang, S. B. (2013). Transformation guided image completion. In *IEEE International Conference on Computational Photography (ICCP)*, pages 1–9.

[52] Hui, Z. and Sankaranarayanan, A. (2016). Shape and spatially-varying reflectance estimation from virtual exemplars. *IEEE transactions on Pattern Analysis and Machine Intelligence*.

[53] Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., et al. (2011). Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *ACM symposium on User interface software and technology*, pages 559–568. ACM.

[54] Jachnik, J., Newcombe, R. A., and Davison, A. J. (2012). Real-time surface light-field capture for augmentation of planar specular surfaces. In *IEEE ISMAR*, pages 91–97.

[55] Jardine, N. and van Rijsbergen, C. J. (1971). The use of hierarchic clustering in information retrieval. *Information storage and retrieval*, 7(5):217–240.

[56] Jarusirisawad, S., Hosokawa, T., and Saito, H. (2010). Diminished reality using plane-sweep algorithm with weakly-calibrated cameras. *Progress in Informatics*, 7:11–20.

[57] Jia-Bin Huang, Sing Bing Kang, N. A. and Kopf, J. (2014). Image completion using planar structure guidance. *ACM Special Interest Group on GRAPHics and Interactive Techniques (SIGGRAPH)*, 33(4):to appear.

[58] Kawai, N., Sato, T., and Yokoya, N. (2013). Diminished reality considering background structures. In *IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 259–260. IEEE.

[59] Kawai, N., Sato, T., and Yokoya, N. (2015). Diminished reality based on image inpainting considering background geometry. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*.

[60] Kawai, N., Sato, T., and Yokoya, N. (2016). Diminished reality based on image inpainting considering background geometry. *IEEE Transactions on Visualization and Computer Graphics*, pages 1236–1247.

[61] Kawai, N., Yamasaki, M., Sato, T., and Yokoya, N. (2012). Ar marker hiding based on image inpainting and reflection of illumination changes. In *IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 293–294. IEEE.

[62] Kim, H., Jin, H., Hadap, S., and Kweon, I. (2013). Specular reflection separation using dark channel prior. In *IEEE CVPR*, pages 1460–1467.

[63] Knill, D. C. and Richards, W. (1996). *Perception as Bayesian inference*. Cambridge University Press.

[64] Kopf, J., Kienzle, W., Drucker, S., and Kang, S. B. (2012). Quality prediction for image completion. *ACM Transactions on Graphics*, 31(6):131.

[65] Korkalo, O., Aittala, M., and Siltanen, S. (2010). Light-weight marker hiding for augmented reality. In *IEEE and ACM International Symposium Mixed and Augmented Reality (ISMAR)*, pages 247–248. IEEE.

[66] Lagger, P. and Fua, P. (2006). Using specularities to recover multiple light sources in the presence of texture. In *IEEE ICPR*, volume 1, pages 587–590.

[67] Lavoué, G. and Mantiuk, R. (2015). Quality assessment in computer graphics. In *Visual Signal Quality Assessment*, pages 243–286. Springer.

[68] Leao, C. W. M., Lima, J. P., Teichrieb, V., Albuquerque, E. S., and Kelner, J. (2011). Altered reality: Augmenting and diminishing reality in real time. In *IEEE Conference on Virtual Reality (VR)*, pages 219–220. IEEE.

[69] Lepetit, V. and Berger, M.-O. (2001). An intuitive tool for outlining objects in video sequences: Applications to augmented and diminished reality. *IEEE and ACM International Symposium Mixed Reality (ISMR)*, 2:159–60.

[70] Li, Q., He, X., Wang, Y., Liu, H., Xu, D., and Guo, F. (2013). Review of spectral imaging technology in biomedical engineering: achievements and challenges. *Journal of biomedical optics*, 18(10):100901.

[71] Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

[72] Lin, J., Seddik, M. E. A., Tamaazousti, M., Tamaazousti, Y., and Bartoli, A. (2019). Deep multi-class adversarial specularity removal. In *Scandinavian Conference on Image Analysis*, pages 3–15. Springer.

[73] Lin, J., Tamaazousti, M., Said, S. H., and Morgand, A. (2017). Color consistency of specular highlights in consumer cameras. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, page 52. ACM.

[74] Lombardi, S. and Nishino, K. (2016). Reflectance and illumination recovery in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):129–141.

[75] Mantiuk, R. K., Tomaszewska, A., and Mantiuk, R. (2012). Comparison of four subjective methods for image quality assessment. In *Computer Graphics Forum*, volume 31, pages 2478–2491.

[76] Marschner, S. R. and Greenberg, D. P. (1997). Inverse lighting for photography. In *Color and Imaging Conference*, volume 1997, pages 262–265. Society for Imaging Science and Technology.

[77] McCamy, C. S., Marcus, H., and Davidson, J. (1976). A color-rendition chart. *J. App. Photog. Eng*, 2(3):95–99.

[78] Meerits, S. and Saito, H. (2015). Real-time diminished reality for dynamic scenes. In *2015 IEEE International Symposium on Mixed and Augmented Reality Workshops*, pages 53–59. IEEE.

[79] Meka, A., Maximov, M., Zollhoefer, M., Chatterjee, A., Seidel, H.-P., Richardt, C., and Theobalt, C. (2018). Lime: Live intrinsic material estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6315–6324.

[80] Mercier, B., Meneveaux, D., and Fournier, A. (2007). A framework for automatically recovering object shape, reflectance and light sources from calibrated images. *International Journal of Computer Vision*, 73(1):77–93.

[81] Morgand, A. (2018). *Un modèle géométrique multi-vues des taches spéculaires basé sur les quadriques avec application en réalité augmentée*. PhD thesis, Clermont Auvergne.

[82] Morgand, A. and Tamaazousti, M. (2014). Generic and real-time detection of specular reflections in images. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 1, pages 274–282. IEEE.

[83] Morgand, A., Tamaazousti, M., and Bartoli, A. (2016). An empirical model for specularity prediction with application to dynamic retuxturing. In *IEEE ISMAR*.

[84] Morgand, A., Tamaazousti, M., and Bartoli, A. (2017). A multiple-view geometric model of specularities on non-planar shapes with application to dynamic retexturing. *IEEE Transactions on Visualization and Computer Graphics*.

[85] Nayar, S. K., Ikeuchi, K., and Kanade, T. (1989). Surface reflection: physical and geometrical perspectives. Technical report.

[86] Newson, A., Almansa, A., Fradet, M., Gousseau, Y., and Pérez, P. (2014). Video inpainting of complex scenes. *SIAM Journal on Imaging Sciences*, 7(4):1993–2019.

[87] Nishino, K., Zhang, Z., and Ikeuchi, K. (2001). Determining reflectance parameters and illumination distribution from a sparse set of images for view-dependent image synthesis. In *IEEE ICCV*, volume 1, pages 599–606.

[88] Patow, G. and Pueyo, X. (2003). A survey of inverse rendering problems. In *Computer graphics forum*, volume 22, pages 663–687.

[89] Phong, B. T. (1975). Illumination for computer generated pictures. *Communications of the ACM*, 18(6):311–317.

[90] Pope, A. and Rees, W. G. (2014). Impact of spatial, spectral, and radiometric properties of multispectral imagers on glacier surface classification. *Remote Sensing of Environment*, 141:1–13.

[91] Poulin, P., Ratib, K., and Jacques, M. (1997). Sketching shadows and highlights to position lights. In *Computer Graphics International, 1997. Proceedings*, pages 56–63. IEEE.

[92] Ramamoorthi, R. and Hanrahan, P. (2001). A signal-processing framework for inverse rendering. In *ACM SIGGRAPH Computer Graphics*, pages 117–128.

[93] Riviere, J., Peers, P., and Ghosh, A. (2016). Mobile surface reflectometry. In *Computer Graphics Forum*, volume 35, pages 191–202.

[94] Sato, I., Sato, Y., and Ikeuchi, K. (1999). Illumination distribution from shadows. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 1, pages 306–312. IEEE.

[95] Shi, J., Dong, Y., Su, H., and Yu, S. X. (2017). Learning non-lambertian object intrinsics across shapenet categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1694.

[96] Siltanen, S. (2015). Diminished reality for augmented reality interior design. *The Visual Computer*, pages 1–16.

[97] Silveira, G. and Malis, E. (2010). Unified direct visual tracking of rigid and deformable surfaces under generic illumination changes in grayscale and color images. *International journal of computer vision*, 89(1):84–105.

[98] Tamaazousti, M. (2013). *L'ajustement de faisceaux contraint comme cadre d'unification des méthodes de localisation: application à la réalité augmentée sur des objets 3D*. PhD thesis.

[99] Tamaazousti, M., Gay-Bellile, V., Collette, S., Bourgeois, S., and Dhome, M. (2011). Nonlinear refinement of structure from motion reconstruction by taking advantage of a partial knowledge of the environment. In *IEEE CVPR*, pages 3073–3080.

[100] Telea, A. (2004). An image inpainting technique based on the fast marching method. *Journal of graphics tools*, 9(1):23–34.

[101] Tiefenbacher, P., Sirch, M., and Rigoll, G. (2016). Mono camera multi-view diminished reality. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8. IEEE.

[102] Torrance, K. E. and Sparrow, E. M. (1967). Theory for off-specular reflection from roughened surfaces. *Journal of the Optical Society of America*, 57(9):1105–1112.

[103] Ward, G. J. (1992). Measuring and modeling anisotropic reflection. *ACM SIGGRAPH Computer Graphics*, 26(2):265–272.

[104] Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

[105] Whelan, T., Salas-Moreno, R. F., Glocker, B., Davison, A. J., and Leutenegger, S. (2016). Elasticfusion: Real-time dense slam and light source estimation. *The International Journal of Robotics Research*, 35(14):1697–1716.

[106] Wiener, N. (1949). *Extrapolation, interpolation, and smoothing of stationary time series*, volume 2. MIT press Cambridge, MA.

[107] Xu, S. and Wallace, A. M. (2008). Recovering surface reflectance and multiple light locations and intensities from image data. *Pattern Recognition Letters*, 29(11):1639–1647.

[108] Yu, Y., Debevec, P., Malik, J., and Hawkins, T. (1999). Inverse global illumination: Recovering reflectance models of real scenes from photographs. pages 215–224.

[109] Zhang, E., Cohen, M. F., and Curless, B. (2016). Emptying, refurnishing, and relighting indoor spaces. *ACM Transactions on Graphics (TOG)*, 35(6):174.

[110] Zhao, W. and Du, S. (2016). Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach. *IEEE Transactions on Geoscience and Remote Sensing*, 54(8):4544–4554.

[111] Ziemer, P. (2013). Design and implementation of a multispectral imaging system. Master's thesis.

Appendix A

Image Completion using Multispectral Imaging

A.1 Introduction

Image completion consists in filling or restoring missing or damaged regions in a visually plausible way. This image processing technique has many applications, such as the removal of unwanted objects in photos and panoramas [64], image restoration [38] and diminished reality [60]. The research in this field has reached an advanced level of maturity, some of the methods being incorporated in raster graphics editors (e.g. PatchMatch [3] in Photoshop CS5[1]). The completion task is non-trivial and is of growing importance in computer vision and computer graphics.

New completion methods were recently proposed to guide the filling of missing regions using prior information about structures [18] and perspectives [50], by using guidance maps [51] or by using statistics of similar patches [46]. This high-level information corresponds to prior knowledge on the geometry of the scene. At last, the completion process is performed and representated with red, green and blue (RGB) values. Rather than employing RGB cameras, multispectral camera-recorders [43] provide more detailed information about the spectrum of objects present in the scene. Those cameras may be of help to address standard computer vision tasks [19], especially when considering the recent introduction of snapshot multispectral camera-recorders [32].

Basically, the content of an image depends on both its geometrical and spectral dimensions [19]. Multispectral images are represented through three dimensional datacubes, where a set of two dimensional images is acquired at different bands of wavelengths using dedi-

[1] http://www.adobe.com/technology/projects/patchmatch.html

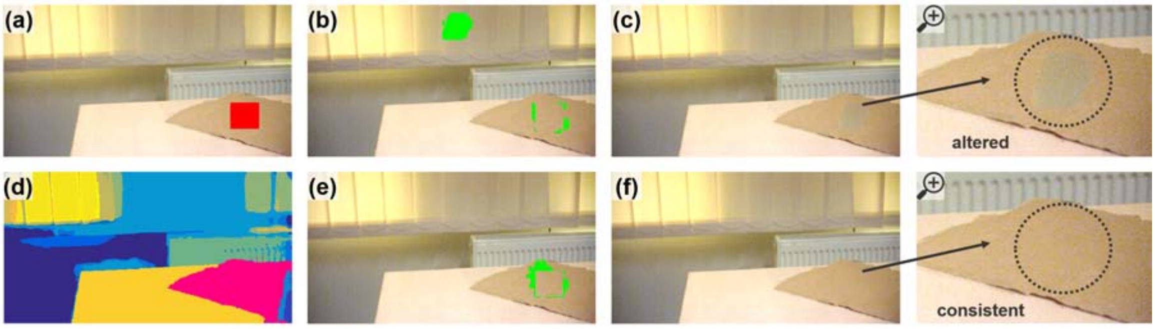


Fig. A.1 Image completion constrained by spectral segmentation. (a) Image recorded by the multispectral camera and converted to RGB. The red box in the paperboard was selected by the user and corresponds to the area to be completed (missing region). (b) Pixels selected by standard exemplar-based completion algorithm (i.e. PatchMatch [3]) to complete the missing region are highlighted in green. The algorithm considers (by mistake) some pixels from the curtains to complete the paperboard, their RGB values being very similar. (c) Resulting completion is visually altered and is partially gray. (d) Spectral segmentation deriving from noise adjusted principal component analysis of the multispectral image. Note that the spectral segmentation produces regions that seem to be consistent with the geometry and materials of the objects. (e) The research is geographically limited to the segments in the neighborhood of the region to be completed (i.e. the magenta segment in figure d). (f) Completion constrained by the spectral segments is more compatible with standard visual assessment in computer vision and computer graphics.

cated optical devices [43]. In the fields of Earth and planetary sciences, datacubes delivered by multispectral or hyperspectral cameras are processed and analyzed to provide relevant information about the chemical composition of the recorded scenes. One of the important advantages of this technique is that physical processes like absorption, reflectance, or fluorescence spectrum can be estimated for each pixel in the image. It allows the detection of chemical changes of objects that cannot be identified with monochromatic or color (RGB) data [70].

The spectral information has been notably employed to characterize ocean color [36], classify glacier surfaces [90] or to sense gypsum on Mars [33]. Also, spectral imaging corresponds to a powerful analytical tool for biological and biomedical research, notably in order to identify tissue abnormalities [70]. The spectral information of a pure material is enough scale-invariant to provide very valuable cues to better understand the contents of an image [19]. Material recognition is presumed to reinforce image processing and understanding techniques such as object detection, object recognition and image segmentation [6].

To date, there have been no studies that analyze the relevance of multispectral imaging in the image completion context. Analyzing multispectral frames instead of RGB frames amounts to process the spectral dimension at each pixel of the image. This information can be used to improve the renderings by properly updating photometric parameters, in particular for diminished reality applications [60].

In this study, we propose to investigate the relevance of multispectral frames applied to image completion, an application initially dedicated to three dimensional RGB images. The study first provides, in section A.2, some basics of multispectral imaging (sensor specifications and pre-processing operations). Because the main purpose consists in better completing images dedicated to visualization, this section also includes elements about the conversion from the recorded multispectral channels to the standard RGB color space.

In section A.3, we describe the behavior of a reference completion algorithm on multispectral datacubes by directly extending its input (from 3 dimensional RGB images to 16 multispectral channels).

Section A.4 presents a better two-step method dedicated to the use of multispectral channels for image completion. A pre-segmentation of the geometry of the scene based on the spectral dimension is described in first step. Research of substitution pixels is then geometrically constrained to a predefined area: only the segments located in the vicinity of the missing region are considered (see figure A.1 for a representative example).

Section A.5 is dedicated to the analysis of results from a perceptual quality assessment procedure based on standard subjective questionnaires over a panel of 20 observers. The proposed method (presented in section A.4) delivers completed images that are more compatible with standard visual assessment in computer vision and computer graphics.

A.2 Multispectral Data

This section presents details about the multispectral device in addition to the image processing operations that were employed to analyze the multispectral data.

A.2.1 Camera Specifications

The multispectral imaging technology we used (figure A.2 a) in this study was designed by IMEC [32]. The device corresponds to a snapshot (i.e. non-scanning) and ultra-compact spectrometer. The camera records the spectral irradiance $I(x,y,\lambda)$ of a scene through a multispectral image, i.e. a 3-D dataset typically called a datacube or hypercube [43]. The

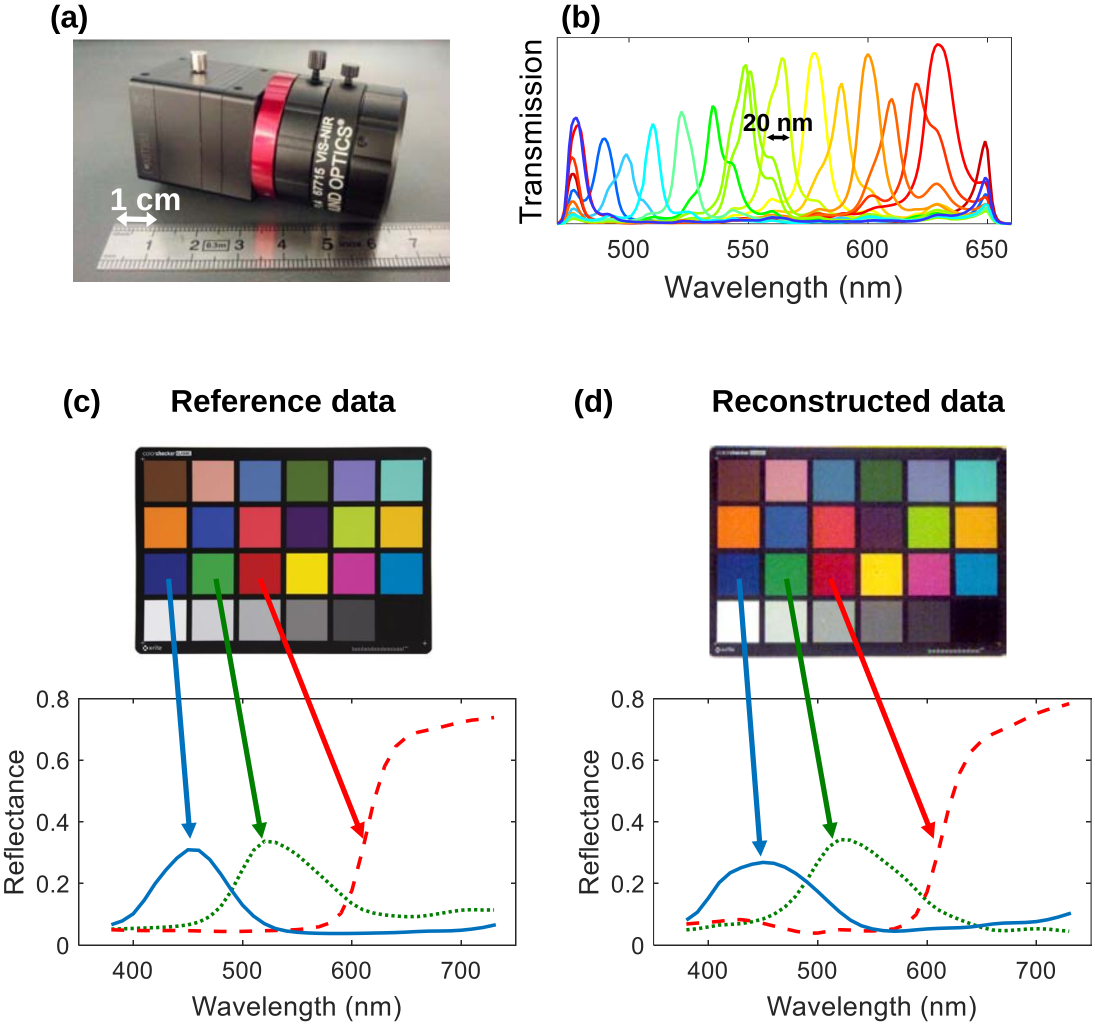


Fig. A.2 Multispectral camera specifications. (a) Snapshot real-time multispectral camera designed by IMEC [32]. (b) Spectral sensitivity of the 16 camera channels, which uniformly encompass most of the visible spectrum (475 to 650 nm). Spectral bandwidth is about 20 nm per channel. In practice, partial correlation between channels results in 14 independent components instead of 16. (c) *ColorChecker Classic* (X-Rite). The color chart contains 24 color patches [77]. Their reference spectra, defined between 380 and 730 nm, are provided by the manufacturer. (d) Image and spectra derived from the multispectral camera. Reference and reconstructed spectra match within up to 90% RMS. The slight discrepancies result from uncertainties in the spectral calibration procedure.

device can nominally deliver 170 datacubes per second in real-time. This value is constrained by the exposure time in practice.

Practically, the camera senses 16 different spectral bands between 475 and 650 nm. The bandwidth of each band is comprised between 15 and 20 nm (figure A.2b). The full resolution of the CMOS sensor is defined to 2048×1024 pixels but reduced to 512×256 pixels for each spectral channel (each cell being formed by a 4×4 multispectral mosaic [32]). Pixel intensity (bit depth) is signed over 10 bits.

A.2.2 Pre-Processing

Spectral Reconstruction

Spectral reconstruction corresponds to a primary procedure essentially employed to calibrate multi or hyperspectral sensors in order to assess apparent reflectances from raw spectral channels [111]. In the present study, spectral reconstruction was performed using a color chart that includes 24 different color patches (see figure A.2c).

Given that all the optical parameters cannot be estimated beforehand, an indirect method was employed to calibrate the multispectral sensor. For the sake of completeness, the interested reader can refer to the original article [111] in order to get the full implementation details. The reconstructed reflectance of the blue, red and green patches of the color chart are illustrated in figure A.2d. The observable discrepancies result from uncertainties on the calibration procedure, which closely depends on the spectral sensitivity responses (figure A.2b).

Multispectral to RGB Conversion

Because completion algorithms deliver images that are displayed on screen and visually evaluated by humans, a conversion to the standard RGB color space is required. In practice, this conversion is achieved using apparent reflectances deriving from the camera calibration procedure (section A.2.2) and by the means of the CIE color matching functions (see figure A.3b). An example of standard RGB conversion is presented in figure A.2d.

A.3 Preliminary Analyses

A.3.1 Significance of the Spectral Sampling

Image completion is based on color and brightness analysis of different image patches. Figure A.3a presents a typical example, where **P1** and **P2** correspond to patches of similar RGB color.

Working with more spectral bands (by increasing the spectral sampling) can be helpful in order to reveal additional relevant information. Figure A.3b presents the spectrum along with the corresponding RGB values of both the **P1** and **P2** patches. Herein, important chromatic

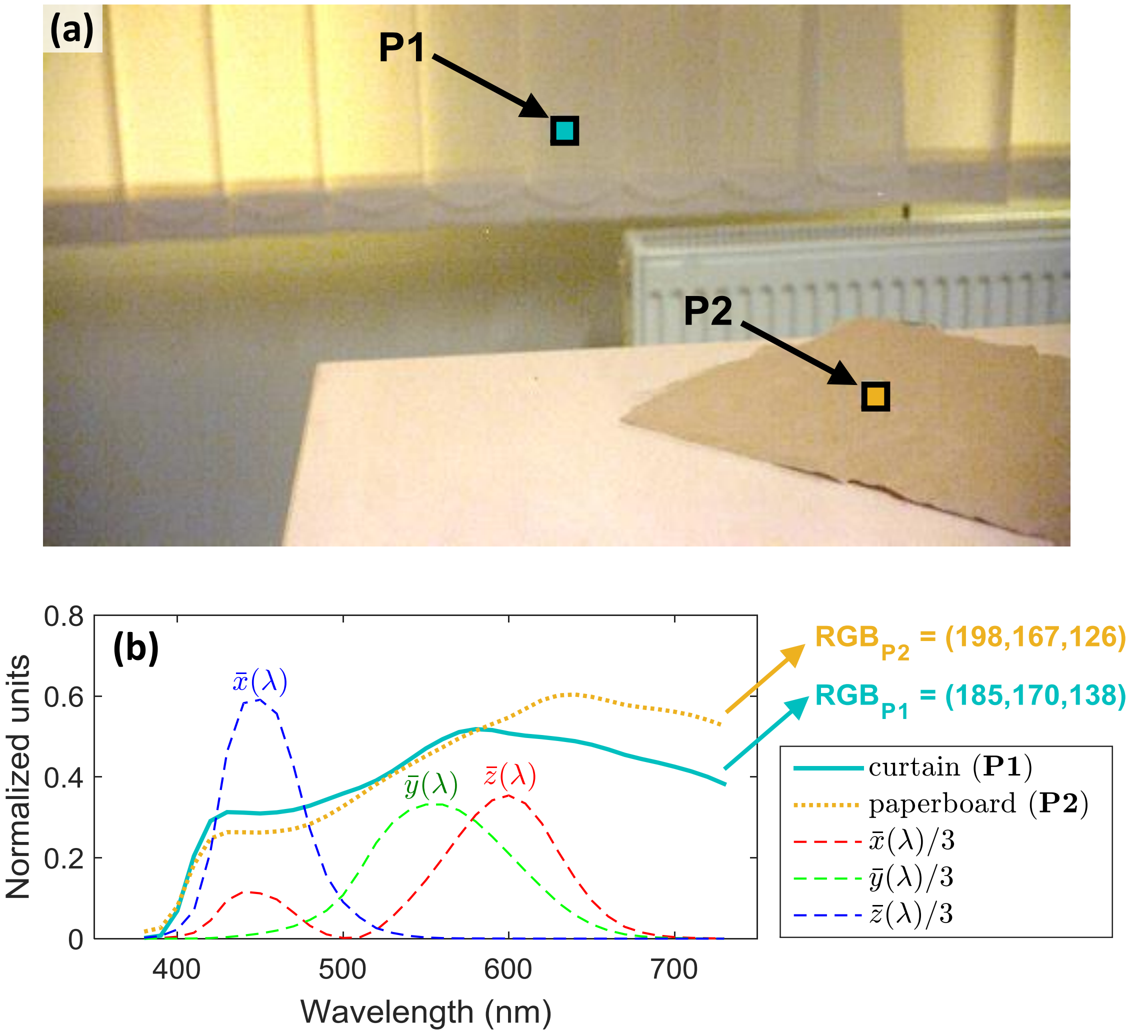


Fig. A.3 Spectral resolution significance. Averaged spectra along with their respective red, green and blue values have been extracted from **P1** and **P2** patches. RGB values indicate that the colors are very similar. Multispectral sampling allows a more precise observation of chromatic differences. Reduction of spectra to three R, G and B values leads to smooth and filter out spectral details, in particular when relevant variations are canceled due to integration by the CIE matching functions ($\bar{x}$, $\bar{y}$ and $\bar{z}$ curves).

differences appear between 590 and 730 nm. These disparities are partially canceled due to integration by the CIE color matching functions ($\bar{x}(\lambda)$, $\bar{y}(\lambda)$ and $\bar{z}(\lambda)$ on figure A.3b). Employing more spectral bands seems relevant in order to better consider chromatic variations when performing image completion.

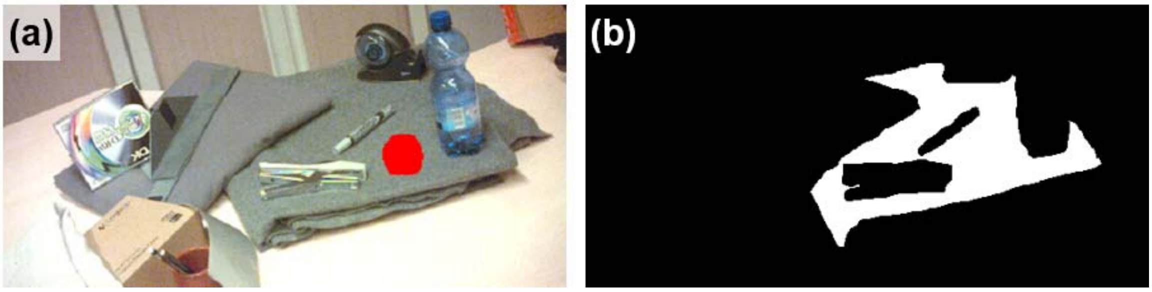


Fig. A.4 Experimental setup. (a) Typical image acquired with the multispectral device and converted to RGB. The red region, which is selected by the user, denotes the area to be completed (missing region). (b) Ground truth material mask (Ω). The white region has been manually segmented and corresponds to the best zone of research (in terms of material) for completion candidates.

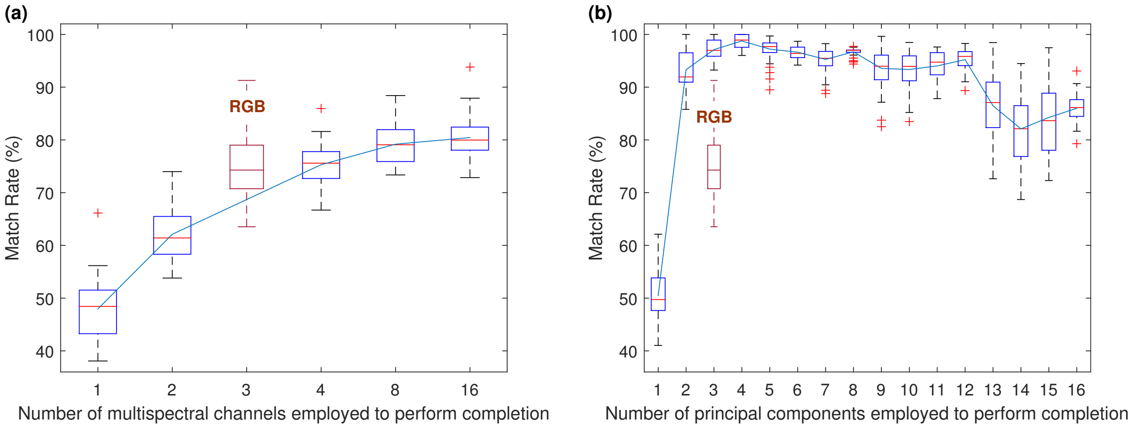


Fig. A.5 Respect of the surrounding materials by the completion algorithm. The match rates are computed between offsets and the ground truth material mask for each image. The presented results integrate all the 500 trials. For comparison purposes, match rates computed using RGB images are indicated on each figure (red boxplot). (a) Match rates computed on raw multispectral channels, starting from single (monochromatic) channel to all the 16 channels. (b) Match rates computed on noise adjusted principal components.

A.3.2 Experimental Procedure

A set composed of 10 different multispectral images was employed to assess the relevance of the multispectral data applied to image completion. The frames were recorded with the multispectral device presented in section A.2.1, the scenes being selected to emphasize current image completion limits. To this purpose, objects and backgrounds of similar colors were employed (figure A.4a).

Each area to be completed was manually chosen and presents no clear gradients and little spatial structure variance. For validation purpose, the regions were defined to avoid entire

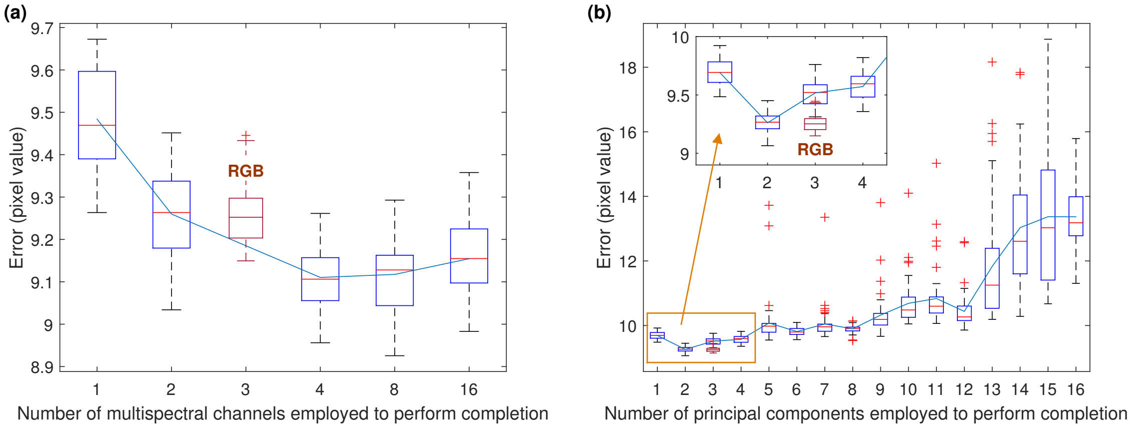


Fig. A.6 Rendering analysis assessed using euclidean errors computed between synthesized and original RGB images. The results are averaged over the 500 trials. For comparison purposes, the errors computed when completion used offset defined on RGB frames are indicated on each figure using a red boxplot. (a) Errors computed using offsets determined on raw multispectral channels. (b) Errors computed using offsets determined on principal components.

overlapping of an object and are comprised on a single material. To evaluate the behavior of the completion procedure, ground truth material masks were manually defined (figure A.4b). They correspond to the region defined by the same material than the one which surrounds the area to be completed. These material masks are also used to evaluate the relevance of the spectral segmentation proposed in this study (see section A.4.2).

We propose to assess the behavior of standard completion algorithm (section A.3.3) in regard to the materials that surround the region to be completed (section A.3.4), in particular when increasing the number of multispectral channels. We also propose to empirically evaluate the quality of the completion by comparing the synthesized area with its original content (section A.3.5).

A.3.3 Implementation Details

PatchMatch [3], which was initially proposed by Barnes et al., is used as a reference image completion technique. The algorithm ensures consistency by solving a global optimization problem and is faster than comparable completion techniques. The method is composed of a sequence of specific steps. The interested reader can refer to the original article [3] in order to get the full implementation details.

Briefly, the method is defined over three main steps: (1) initialization: a random patch offset is given to each pixel at the coarsest pyramid level of the image. The result is propagated to the next pyramid level where a propagation and random search steps are applied at each level; (2) propagation: the pertinence of the offsets is evaluated with respect to the neighboring patches at each iteration using an objective function; (3) random search: a search step is employed to look for better patch within a concentric radius around the current offset. The new offset is adopted if the new objective function is lower.

A particular implementation of the initialization step was employed in this study. A first exhaustive search of the best matching offsets is performed [47] instead of a random one. Also, the patch size has been set to 13×13 pixels. Because of the random process included in PatchMatch, 50 trials per image were launched to compute statistical tendencies, a single run being non representative.

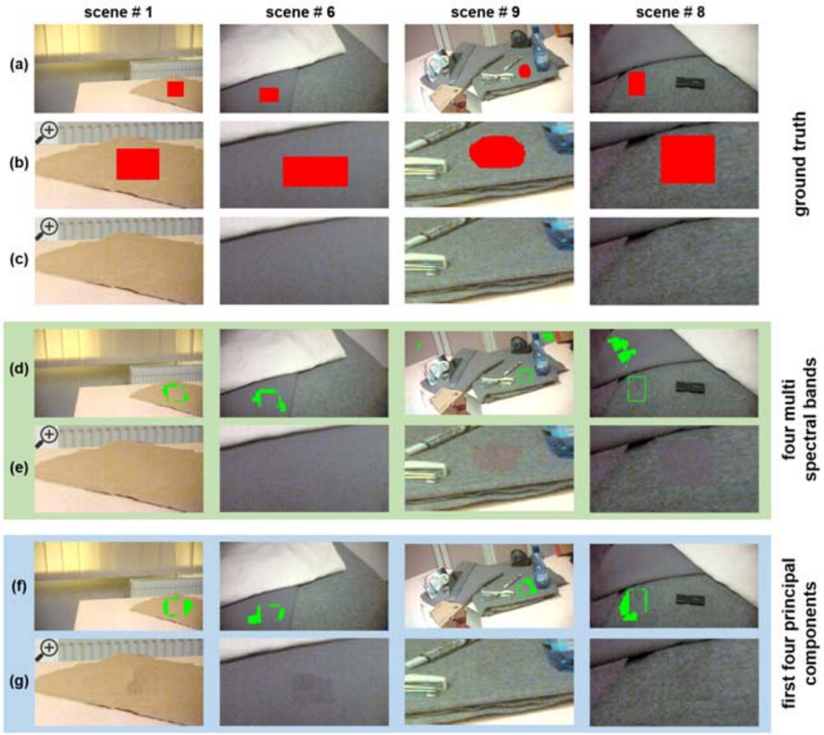


Fig. A.7 Spectral completion. (a) Source image with (b) its corresponding close-up view. The red pattern corresponds to the area to be completed. (c) Ground truth (close-up). (d) Pixels selected using four multispectral channels to complete the missing region are highlighted in green. (e) Completion results (close-up) based on the selected pixels from (d). (f) Pixels selected using the first four principal components to complete the missing region are highlighted in green. (g) Completion results (close-up) based on the selected pixels from (f).

A.3.4 Materials Consideration

In this section, we propose to assess the behavior of the completion algorithm in regard to the materials that surround the region to be completed, in particular when increasing the number of multispectral channels.

The full image I is separated into two disjoint sets: T corresponds to the target (or missing) region, completed using pixels in S (source region). $I = T \cup S$, $T \cap S = \varnothing$ and $S \neq \varnothing$. The image completion algorithm replaces all pixels included in T.

The offsets represent the difference of position between a pixel in the area to be completed (target region) and its corresponding candidate in the source region. Offsets are defined with a mapping function f that maps each target position $p \in T$ to a source position $q \in S$ (see figure A.1b and A.1e for typical examples):

$$f : T \to S \tag{A.1}$$

f corresponds to a transformation that solves a global minimization problem and is determined for each target pixel. The synthesized image is then created by replacing all target pixels with their corresponding source pixels. It is important to note that only the offsets, i.e. the difference of position between a pixel included in the area to be completed and its corresponding candidate in the rest of the image, are susceptible to fluctuate. The synthesizing procedure (pixel copy) is ultimately performed on RGB frames using the defined offsets.

To understand if the completion algorithm is able to correctly use pixels from surrounding materials, the percentage of good match (α in eq. A.2) between the offsets and the ground truth material mask was assessed for each of the 10 input images. It corresponds to the number of times the completion algorithm uses a pixel from the ground truth material region over the total number of pixels in the target area:

$$\alpha = \frac{100}{N} \sum_{i=1}^{N} \gamma(p_i) \tag{A.2}$$

$$\gamma(p) = \begin{cases} 1, & f \in \Omega \\ 0, & \text{else} \end{cases} \tag{A.3}$$

Where Ω corresponds to the ground truth material region (figure A.4b). γ is defined for each target pixel (p). N corresponds to the total number of pixels from the target region and α to the match rate (units: %).

Results are presented in figures A.5a and A.5b using boxplot representations. Each boxplot includes 500 computed match rates (10 images recorded by the multispectral camera × 50 completion trials per image). For comparison purposes, the match rates computed using RGB images were reported on these figures (red boxes).

Figure A.5a presents the match rates computed when completion is performed on raw multispectral channels. Starting from all the 16 channels, we progressively averaged the spectral image two channels by two channels until reaching a single channel (monochromatic image). Figure A.5b presents the same percentage of good match, but when performing completion on principal components. The latter were computed from a noise adjusted principal component analysis, a transformation developed to sort principal components by image quality (decreasing image quality with increasing component number). We have employed minimum/maximum autocorrelation factors to estimate the noise covariance matrix. The method has been proposed by Green et al. [37] and uses between-neighbor differences to estimate the noise covariance.

Results presented in figure A.5a exhibit an increase of the match rates that are correlated with the augmentation of the number of channels. Also, the boxplots length indicates that the variance tends to simultaneously decrease. Adding a more precise spectral information to the completion algorithm leads to better considerate the physical properties of materials. Subtle variations that were not necessarily observable in the standard RGB color space are considered (see section A.3.1).

Image completion based on principal components (figure A.5b) tends to better consider the surrounding materials, the maximum median value being equal to 99% (instead of maximally 80% when considering raw multispectral channels). In addition, only four components are necessary to achieve this score. The last principal components containing more and more noise, the induced artifacts generate a bias that leads the completion to pick patches in a random fashion, thus reducing the mean percentage of good match while increasing the variance.

A.3.5 Rendering Analysis

In this section, we propose to empirically assess the quality of the completion by comparing the synthesized area with its original content using an error function. The latter corresponds to the euclidean distance based on the R, G and B channels and is computed for each pixel of the target region.

Figure A.6a presents the euclidean errors computed using offsets that where determined on raw multispectral channels. As before (see section A.3.4) and starting from all the 16 channels, we progressively averaged the spectral image two channels by two channels

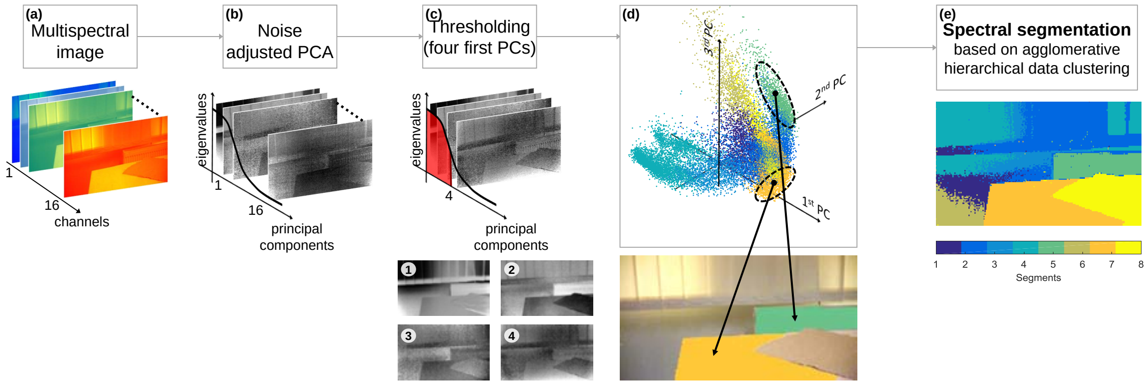


Fig. A.8 Spectral segmentation procedure. (a) Raw multispectral image defined over 16 spectral channels. (b) Noise adjusted principal component analysis [37]. Eigenvectors correspond to images with decreasing eigenvalues. (c) Thresholding operation is applied to recover the first four principal components, which are more adapted to the context (see section A.3.4). (d) Projection of the pixel values in the spectral space defined by the first four principal components (only the first three dimensions are represented). In this spectral space, clusters (see dashed-line ellipses) match with the geometry and materials of the scene. (e) Resulting segmentation, based on hierarchical data clustering (number of clusters: 8).

until reaching a single channel (monochromatic image). Figure A.6b presents the same information, but when performing completion on principal components. A close-up view is displayed on the top of the figure to identify the error minimum. Errors computed when completion is performed on standard RGB images are respectively reported on the two figures using red boxes.

Completion based on four multispectral channels (figure A.6a) presents the general minimum error. From figure A.6b, completion based on the first two principal components presents the minimum error. Employing more principal components gives worse completion results. This effect is inherent to the noise adjusted principal component transform: the last components containing more and more noise, the induced artifacts generate a bias that leads the completion to pick patches in a random fashion.

In addition to these statistical tendencies, illustrative completion results are presented in figure A.7 to visually compare renderings. From these results, we can conclude that completion based on four multispectral channels produces plausible results when the materials are respected (i.e. when only the pixels included in the region defined by the material that surrounds the missing region are used for completion. See figure A.7, images # 1 and 6, for a typical example). In comparison, completion based on the first four principal components produces less consistent results. The chromaticity (colors) is respected but the intensity (brightness) seems inconsistently distributed. In contrast, completion based on four

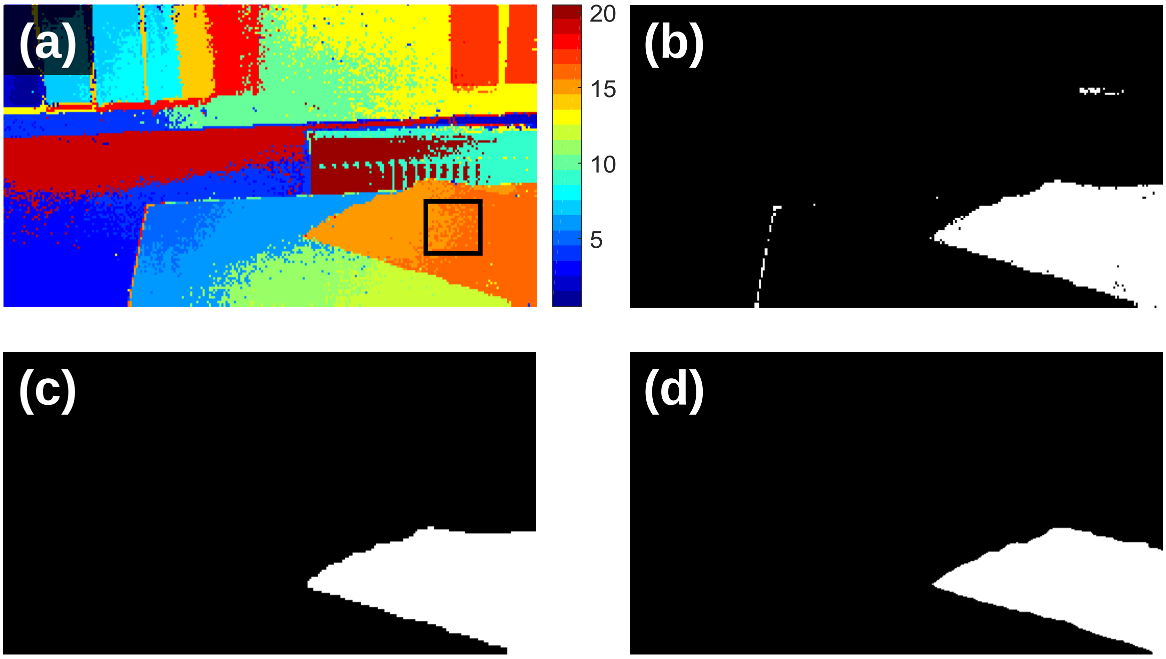


Fig. A.9 Segments merging and post-processing treatments. (a) Spectral segmentation based on agglomerative hierarchical data clustering (number of clusters: 20). The black rectangle indicates the area to be completed. (b) The segments located in the vicinity of the region to be completed are merged to form a single, binary mask. (c) Post-processing treatments are applied to remove small group of pixels and fill small holes. (d) Ground truth material mask (manually segmented).

multispectral channels tends to produce chromatic inconsistencies when the materials are not respected (see results of images # 8 and 9 on figure A.7). This time, completion based on the first four principal components delivers more plausible results, even if brightness discrepancies can still be noted.

A.4 Image Completion Constrained by Spectral Segmentation

A.4.1 Motivation

Two important points emerged from the preliminary analyses results (sections A.3.4 and A.3.5):

- Noise adjusted principal components, computed from the full spectral data, tend to better consider the materials of the scene: this particular representation constitutes a good way to separate pixels from different material classes and therefore ensure the stability of

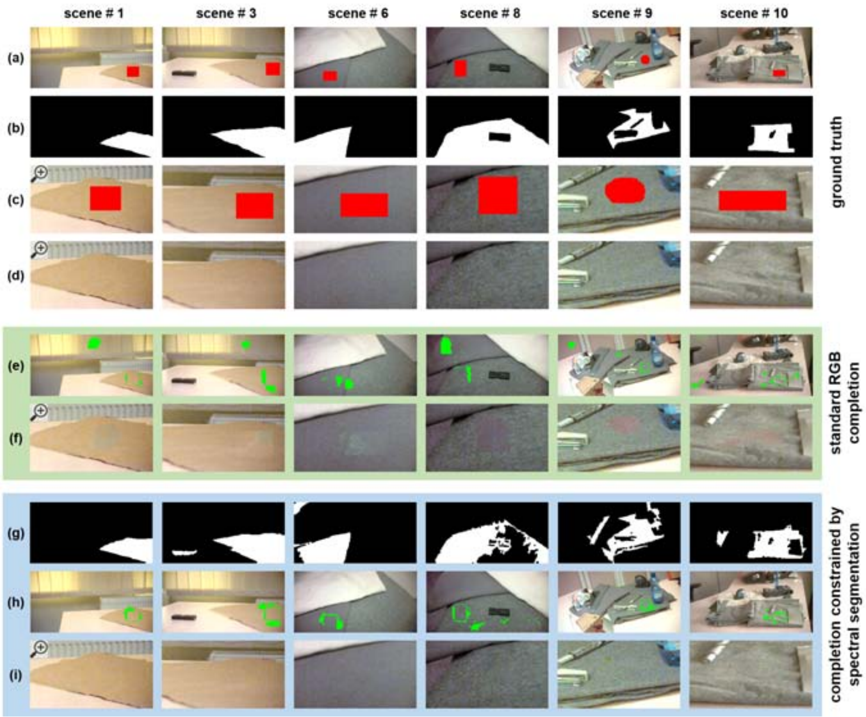


Fig. A.10 Comparisons with representative baseline algorithm [3]. (a) Source image. (b) Ground truth material mask. (c) Corresponding close-up view. In (a) and (c), the red pattern indicates the area to be completed. (d) Ground truth (close-up). (e) Offsets computed using baseline completion method (green pixels). They correspond to the pixels used to complete the missing region. (f) Completion results based on the selected pixels from (e). (g) Spectral segmentation mask. Research of substitution pixels is geometrically constrained to the white area. (h) Offsets computed using the method proposed in this study (green pixels). (i) Completion results based on the selected pixels from (h).

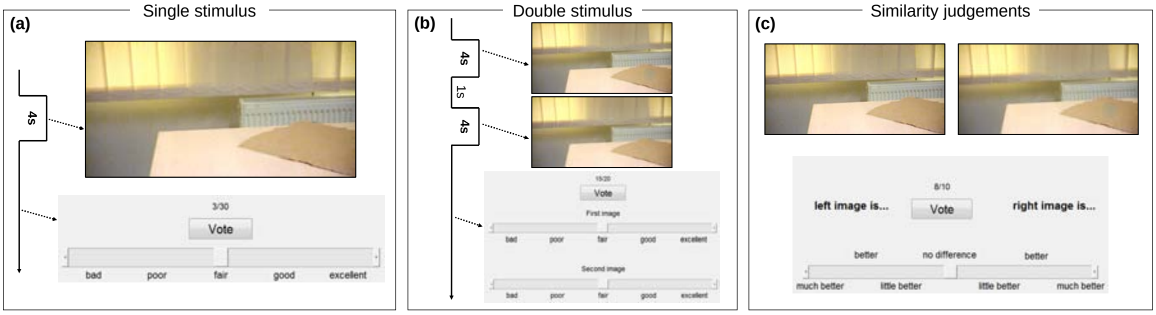


Fig. A.11 Overview of the three subjective methods employed in this study to assess image quality. (a) Single stimulus. Observers have to rate the quality of the displayed image. (b) Double stimulus. Observers must rate the quality of the first and the second image. (c) Similarity judgments. Observers have to express their preference by evaluating the quality differences between the two displayed images.

the completion in terms of materials. The first four principal components gives a maximum match rate (figure A.5b).

- Completion based on multispectral channels produces more plausible results than completion based on principal components when the materials are respected (figure A.7). Completion based on four multispectral channels presents the minimum error (figure A.6a).

Thereby, we can conclude that principal components must be considered in order to fill the missing region with pixels included in the same material area. In addition, the completion must be based on the raw multispectral channels to ensure a consistent and plausible rendering. The method we propose in this section is based on these two observations: completion is performed on four raw spectral channels but limited to a predefined and coherent area. The latter is estimated through spectral segmentation based on the first four principal components.

A.4.2 Spectral Segmentation

Method

The full pipeline is presented graphically in figure A.8. The input of the method corresponds to the raw multispectral image, where each pixel is defined by its 16 points spectral signature (figure A.8a). Noise adjusted principal component transform [37] is computed (figure A.8b) to reduce the input dimensionality and, based on the results presented in section A.3.4, only the first four components are retained for further processing (figure A.8c).

The spectral segmentation is based on agglomerative hierarchical clustering, which consists in grouping data by creating a cluster tree (*dendrogram*). The similarity between every pair of pixels is firstly evaluated by computing euclidean distances. Note that each pixel is defined by four different coordinates, one coordinate by principal component. The distance information is used to link pairs of pixels that are close together into binary clusters. Each binary cluster is made up of two pixels. The newly formed clusters are then linked once again to create bigger clusters using the Ward's method (minimum increase of sum-of-squares)

Scene #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**All scenes**	
mean	**All scenes**											
standard deviation height**Precision (%)**	92	88	90	87	91	89	96	88	58	72	**85**	**11**
Recall (%)	100	100	96	97	36	91	98	63	89	95	**86**	**21**
Jaccard's distance (%)	92	88	87	85	35	82	94	58	54	70	**74**	**20**

Table A.1 Evaluation of the spectral segmentation proposed in this study (see section A.4.2). Precision and recall indexes, as well as Jaccard's distance, were computed using ground truth material masks (figure A.4).

[104]. This step is repeated until all the pixels in the original data set are linked together, thus forming a hierarchical tree.

The tree may inherently separate the data into distinct clusters, in particular for dendrograms created from groups of densely packed pixels. These groups may correspond to pixels of similar materials. For example, if we only consider the second principal component (presented in figure A.8c) the hierarchical tree will contain four large and separate clusters: the paperboard (dark pixels), the table (dark-gray pixels), the radiator (white pixels) and the background (light-gray pixels). The hierarchical cluster tree is pruned to partition the data set into separated clusters. Usually, the number of clusters must be carefully selected to avoid over- and under-segmentation. Under-segmentation is not permitted: pixels that belong to different materials will be grouped in a single segment, thus resulting in a probable inaccurate completion. The segments located in the vicinity of the region to be completed being merged (figure A.9), over-segmentation is tolerated. To properly perform completion, the fused segments of interest (figure A.9c) must include an acceptable amount of pixels.

Practically, the function *clusterdata* included in Matlab (The MathWorks Inc.) was employed. Euclidean distance and Ward's method were used to respectively compute every distance and create the hierarchical tree. As presented in figure A.9, post-processing treatments were developed to remove artifacts. In particular, morphological operations were employed to remove small isolated groups of pixels (surface area $\leqslant$ 200 pixels) and fill small holes (morphological closing using a disk-shaped structural element of radius 3 pixels).

Evaluation

The spectral segmentation is evaluated through Jaccard's distance and precision and recall indexes [55]. The metrics were computed between the spectral segmentation, given by its binary mask (figure A.9c), and the ground truth material mask (figure A.9d) for each scene. The results are presented in table A.1. Generally, the average values for the precision and recall indexes are higher than 85%.

A.4.3 Constrained Multispectral Completion

The binary mask delivered by the spectral segmentation procedure (figure A.9c) is employed to constrain completion in a predefined region. Technically, we deactivate the research process on source pixels located outside the segmentation area.

Note that completion is constrained by spectral segmentation, which is based on the first four principal components (section A.3.4), but ultimately performed by analyzing pixel values on four multispectral channels (section A.3.5). Some excerpts are illustrated in figure

A.10. All the synthesized images were compared against baseline (standard RGB completion) through subjective quality assessment metrics.

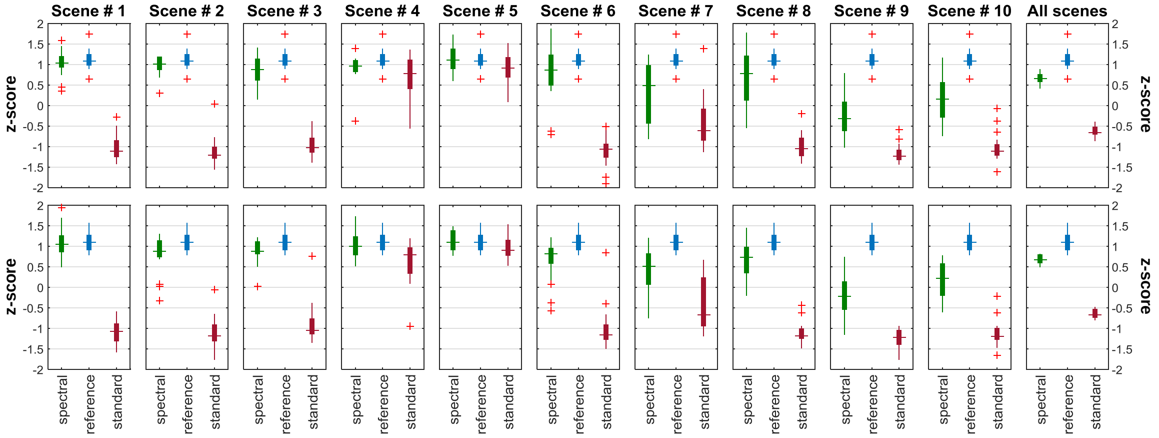


Fig. A.12 Ratings for each scene. Figures exhibit z-scores from single stimulus (first row) and double stimulus (second row). On each figure, the left boxplot has been formed using ratings from images completed by the proposed method (completion constrained by spectral segmentation, see section A.4). The blue central boxplot corresponds to ratings from reference (unmodified) images and the red boxplot on the right to ratings from images completed by baseline method (standard RGB completion). Each boxplot integrates ratings results over all observers, the central mark corresponding to the median, the edges of the box to the 25th and 75th percentiles and the whiskers to the most extreme data points (not considered outliers). Outliers are plotted individually using red crosses.

A.5 Perceptual Quality Assessment

A.5.1 Introduction

Most of computer graphics rendering methods require perceptually plausible results: simple pixel intensity error computed between synthesized and original images (see section A.3.5) does not necessarily reflect and guarantee perceived image quality [67]. Image quality assessment consists in providing a metric that expresses overall quality by rating and ranking methods. Image and video quality assessment has been particularly employed in video compression and transmission applications [75].

To assess visual quality as perceived by observers, ratings and preferences are recorded through subjective questionnaires. These two metrics have been widely used in experimental sciences to assess relative judgments from human participants [67]. Decision times, which are related to the degree of difficulty encountered by the observers to perform the tasks, were also recorded. Quality is assessed for each of the 10 scenes recorded by the multispectral camera (see section A.3.2). The experiment was conducted by 20 different observers (17 males and 3 females, 25–37 years). Three images are employed for each scene to perform quality assessment: (1) image completed by standard RGB method; (2) image completed by the technique proposed in this study (spectral completion); and (3) reference (unmodified) image.

A.5.2 Assessment Methods

All observers received a prior explanation before the beginning of the session. The experiment started with a short training session, in which observers could manipulate the interface and perform training tasks. To avoid effects caused by side variables, all test sets were presented to each observer in a random fashion. Also and to avoid fatigue, no session took longer than 20 minutes. A typical session lasts approximately 15 minutes.

Three methods were employed to assess image quality: (1) *single* and (2) *double stimulus* represent continuous rating while (3) *pairwise similarity judgment* method is employed to evaluate relative preference between two images [75].

Single stimulus: observers judge the quality on a continuous 5-point Likert scale [71]. Each image is displayed for only four seconds. After that short period, a voting interface is displayed on screen. Five categories are indicated right over the continuous scale: bad, poor, fair, good and excellent (figure A.11a). Reference images are included into the set. Thus, observers must evaluate a set composed of 30 randomly arranged images, which includes 10 reference images and 20 completed images (10 by standard RGB completion and 10 by the spectral completion proposed in this study).

Double stimulus is similar to the single stimulus method, except that a reference and a completed image are successively displayed in random order one after another, each one for four seconds (figure A.11b). Observers are asked to independently evaluate the quality of the first and the second image. Herein, observers rate 40 images (20 related to RGB completion and 20 related to spectral completion).

Pairwise similarity judgment: observers are asked to mark their preference by indicating how large the difference in quality is between images synthesized by each of the two completion methods (figure A.11c). A continuous 7-point scale has been employed. Observers can select the central position if no differences were identified between the pair of images.

A.5.3 Results and Analysis

Rating Methods. It has been shown that direct rating results correspond to very unreliable estimates [75]. Thus, we choose to present only differential scores in this section. The latter were computed between pairs of images, in particular between reference and completed images and by means of difference mean opinion scores (eq. A.4):

$$d_{i,j,k} = r_{i,\mathrm{ref}(k),k} - r_{i,j,k} \tag{A.4}$$

$$z_{i,j,k} = \frac{d_{i,j,k} - \overline{d_i}}{\sigma_i} \tag{A.5}$$

$r_{i,j,k}$ corresponds to the rating for a given (reference or completed) image. Indexes correspond to i-th observer, j-th completion method (standard or spectral) and k-th scene. ref(k) corresponds to the reference for scene k. z-scores (eq. A.5) are computed to adjust scale variations between observers in order to properly compare results. To unify scales, a common way consists in normalizing opinion scores by removing the mean ($\overline{d_i}$ in eq. A.5) and unifying standard deviation across observers (σ_i in eq. A.5).

Results for both single and double stimulus experiments are presented in figure A.12. Generally and for both experiments, observers showed preference for images completed by the method proposed in this study: the z-scores are significantly higher than those computed from images completed by the baseline method. In addition, we can notice that observers gave similar opinions for scenes # 4 and 5, indicating that both completion methods performed identically, and particularly well, on these two scenes.

Pairwise similarity judgment. In order to be compared and because each observer could employ a different range of voting values, quality judgments are normalized per observer: each vote has been divided by the observer global standard deviation. The results is similar to z-scores, except that the mean value ($\overline{d_i}$) is not removed. Similarity judgments include preferences, the sign indicating which image was judged better.

Results are presented in figure A.13a. Positive values indicate that images completed by the method we propose in this study (spectral completion) is preferred to images completed by the baseline method. Results produced by spectral completion was preferred in 186 over a total of 200 votes, which correspond to 93% of the total number of votes. 7 votes over 200 (3.5% of the total number) indicate no difference in quality between the two completion methods (values located on the zero axis in figure A.13a). Finally, 7 votes over 200 (3.5%) were in favor of images completed by the baseline method (negative values in figure A.13a).

From these results, we can conclude that the quality of images produced by the spectral completion was preferred in most cases. Except for scenes # 4 and 5, all judgments opt in favor of the method we propose in this study. In accordance with results from single and double stimulus experiments (figure A.12), both completion methods performed well for these two particular scenes, their ratings being close to the reference image. Thus, it appears that observers were less able to distinguish differences in quality between the images produced by the two completion methods. This observation seems to be consistent with the time took by the observers to complete the experiment (figure A.13b), which is significantly higher for scenes # 4 and 5 than for the other scenes.

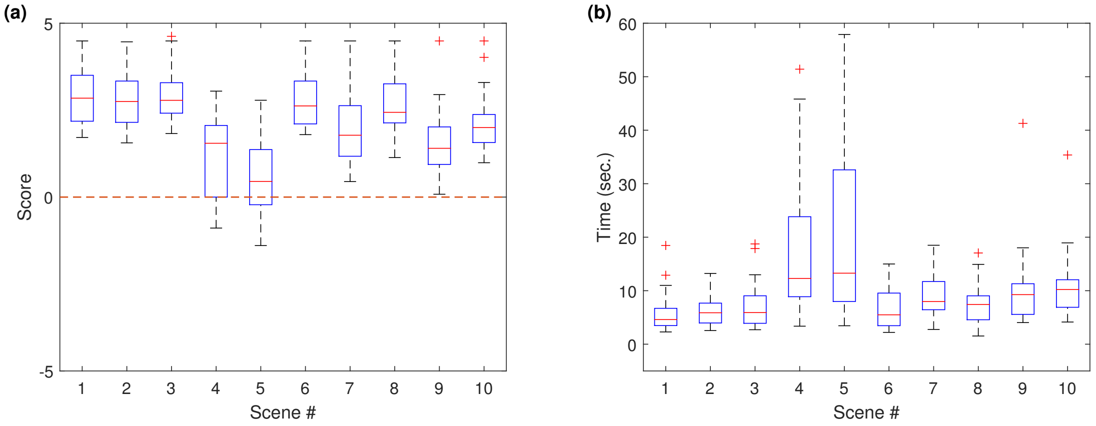


Fig. A.13 Pairwise judgments and time needed to complete the experiment. (a) Pairwise judgments for each scene. Results are presented in normalized units: each vote has been divided by the observer global standard deviation. Positive scores indicate that images completed by the method we propose in this study (spectral completion) is preferred to images completed by the baseline method (standard RGB completion). (b) Time needed to complete the pairwise comparison experiment.

A.6 Discussion

Employing snapshot multispectral cameras instead of hyperspectral ones ensures a real-time exploitation of the method, which corresponds to a necessary prerequisite for many practical applications. In contrast, the direct integration of the spectrum, signed over 16 different values, imposes a drastic extension of computational times. This limitation was considered by integrating dimensionality reduction transforms to the method: based on preliminary analyses (sections A.3.4 and A.3.5), the first four principal components were used to segment the scene while four spectral channels were employed to perform completion by determining which pixels must be copied into the missing region.

Incorporating recent completion techniques (e.g. constraining the completion process with guidance maps [51] or using statistics of similar patches [46]) to the method proposed in this study could be relevant and of interest but is out of the scope of this work. Indeed, the main objective of this study consists in comparing completion based on multispectral images against completion based on RGB images. Adding supplementary constraints, like prior information about structures or guidance maps, may tend to denaturate this comparison.

A.6.1 Limitations

Improvement of the database. The multispectral database currently includes 10 multispectral indoor scenes. The latter were selected to emphasize current image completion limits. To this purpose, objects and backgrounds slightly textured and of similar color were employed. Due to the random process included in PatchMatch, 50 trials per image were launched to compute statistics. Despite the low number of images included in the database, we believe that the tendencies presented in sections A.3.4 and A.3.5 are adequately representative.

Spatial-spectral clustering. The spectral segmentation developed in this study (section A.4) is based on research of clusters in the spectral space. They tend to respect the geometry of the objects but no explicit information, like material-invariant features such as shape or texture for example, is currently incorporated into the method.

A.6.2 Future Works

In regard to the limitations exposed beforehand, the first milestone will consist in expanding the database by including varied indoor and natural scenes.

Developments will be conducted to improve spectral segmentation by coupling spatial and spectral dimensions using Schrödinger Eigenmaps [17], a recent technique that extends Laplacian Eigenmaps in order to fuse spatial and spectral information through nondiagonal potentials. Also, deep learning [110] and support tensor machine [39] correspond to promising avenues that need to be examined.

A.7 Conclusion

We have proposed, in this study, to assess potential of multispectral imaging applied to image completion. Regions to be completed were chosen to present no clear gradients and slight textures. This lack of variance in the spatial structure coupled to the presence of objects of similar color within the image leads to repeated RGB completion failures. Herein, the contribution of the spectral information is of interest and allows better discrimination and, therefore, an increasing rate of successful completion.

Preliminary results indicate that direct exploitation of completion algorithms by extension of the spectral channels shows only minimum enhancement. Based on these observations, we proposed a two-step method dedicated to the use of multispectral channels for image completion. A pre-segmentation of the scene has been developed to geometrically constrained the research of substitution pixels to a predefined area. Only the segments located in the vicinity of the missing region are considered. Results indicate that image completion constrained by spectral segmentation improves rendering consistency and simultaneously ensures better stability in terms of materials.

Results were validated using numerical criteria and perceptual assessment experiments. The proposed method delivers completed images that are more compatible with standard visual assessment in computer vision and computer graphics. Snapshot multispectral devices correspond to breakthrough technologies that can be employed to improve computer vision methods by accurately sensing the physical properties of a scene. This study shows for the first time the potential of snapshot multispectral imaging applied to computer vision and particularly to image completion.