Chapter 9
Physical Integrity

Christian Riess

Physics-based methods anchor the forensic analysis in the physical laws of image and video formation. The analysis is typically based on simplifying assumptions to make the forensic analysis tractable. In scenes that satisfy such assumptions, different types of forensic analysis can be performed. The two most widely used applications are the detection of content repurposing and content splicing. Physics-based methods expose such cases with assumptions about the interaction of light and objects, and about the geometric mapping of light and objects onto the image sensor.

In this chapter, we review the major lines of research on physics-based methods. The approaches are categorized as geometric and photometric, and combinations of both. We also discuss the strengths and limitations of these methods, including an interesting unique property: most physics-based methods are quite robust to low-quality material, and can even be applied to analog photographs. The chapter closes with an outlook and with links to related forensic techniques such as the analysis of physiological signals.

9.1 Introduction

Consider a story that is on the internet. Let us also assume that an important part of that story is a picture or video to document the story. If this story goes viral, it becomes potentially relevant also to classical journalists. Relevance may either inherently be given, because the story reports a controversial event of public interest, or relevance may emerge from the sole fact that a significant number of people are discussing a

C. Riess (✉)
IT Security Infrastructures Lab, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany
e-mail: christian.riess@fau.de

H. T. Sencar et al. (eds.), *Multimedia Forensics*, Advances in Computer Vision and Pattern Recognition, https://doi.org/10.1007/978-981-16-7621-5_9

particular topic. Additionally, quality journalism also requires confidence in the truth value that is associated with a news story. Hence, classical journalistic work involves the work of eye witnesses and other trusted sources of documentation. However, stories that emerge from internet sources are oftentimes considerably more difficult to verify for a number of reasons. One reason may be because the origin of the story is difficult to determine. Another reason may be that reports are difficult to verify because involved locations and people are inaccessible, as it is oftentimes the case for reports from civil war zones. In these cases, a new specialization of journalistic research formed in the past years, namely journalistic fact checking. The techniques of fact checkers will be introduced in Sect. 9.1.1. We will note that the class of physics-based methods in multimedia forensics is closely related to these techniques. More precisely, physics-based approaches can be seen as computational support tools that smoothly integrate into the journalistic fact checking toolbox, as highlighted in Sect. 9.1.2. We close this section with an outline for the remaining chapter in Sect. 9.1.3.

9.1.1 Journalistic Fact Checking

Classical journalistic research about a potential news story aims to answer the so-called "Five W" questions who, what, why, when, and where about an event, i.e., who participated in the event, what happened, why, and at what time and place it happened. The answers to these questions are typically derived and verified from contextual information, witnesses, and supporting documents. Investigative journalists foster a broad network of contacts and sources to verify such events. However, for stories that emerge on the internet, this classical approach can be of limited effectivity, particularly if the sources are anonymized or covered by social media filters. Journalistic fact checkers still heavily rely on classical techniques, and, for example, investigate the surroundings in social networks to learn about possible political alignments and actors from where a message may have originated. For images or videos, context is also helpful.

One common issue with such multimedia content is repurposing. This means that the content itself is authentic, but it has been acquired at a different time or a different place than what is claimed in the associated story. To find cases of repurposing, the first step is a reverse image search as it is possible with specialized search engines like `tineye.com` or `images.google.com`.

Furthermore, a number of open-source intelligence tools are available for a further in-depth verification of the image content. Depending on the shown scene and the context of the news story, photogrammetric methods allow to validate time and place from the position of the sun and known or estimated shadow lengths of buildings, vehicles, or other known objects. Any landmarks in the scene, written text, or signs can be further used to constrain the possible location of acquisition, combined with, e.g., regionally annotated satellite maps. Another type of open-source information is monitoring tools for flights and shipping lines. Such investigations require a high

manual work effort to collect and organize different pieces of evidence and to assess their individual credibility. In journalistic practice, however, these approaches are currently the gold standard to verify multimedia content from sources with very limited contextual information. Unfortunately, this manual effort also implies that an investigation may take several days or weeks. While it would oftentimes be desirable to quell wrong viral stories early on, this is in many cases not possible with current investigative tools and techniques.

9.1.2 Physics-Based Methods in Multimedia Forensics

In multimedia forensics, the group of physics-based methods follows the same spirit as journalistic fact checking. Like the journalistic methods, they operate on the content of the scene. They also use properties of known objects, and some amount of outside knowledge, such as the presence of a single dominant light source. However, there are two main differences. First, physics-based methods in multimedia forensics operate almost exclusively on the scene content. They use much less contextual knowledge, since this contextual information is strongly dependent on the actual case, which makes it difficult to address in general-purpose algorithms. Second, journalistic methods focus on validating time, place, and actors, while physics-based methods in multimedia forensics aim to verify that the image is authentic, i.e., that the scene content is consistent with the laws of physics.

The second difference makes physics-based forensic algorithms complementary to journalistic methods. They can add specific evidence whether an image is a composite from multiple sources. For example, if an image shows two persons on a free field who are only illuminated by the sun, then one can expect that both people are illuminated from the same direction. More subtle cues can be derived from the laws of perspective projection. This applies to all common acquisition devices, since it is the foundation of imaging with a pinhole camera. The way how light is geometrically mapped onto the sensor is tied to the camera and to the location of objects in the scene. In several cases, the parameters for this mapping can be calculated from specific objects. Two objects in an image with inconsistent parameters can indicate an image composition. Similarly, it is possible to compare the color of incident light on different objects, or the color of ambient light in shadow areas.

To achieve this type of analysis, physics-based algorithms use methods from related research fields, most notably from computer vision and photometry. Typically, the analysis analytically solves an assumed physical model for a quantity of interest. This has several important consequences, which distinguish physics-based approaches from statistical approaches in multimedia forensics. First, physics-based methods typically require an analyst to validate the model assumptions, and to perform a limited amount of annotations in the scene to access the known variables. In contrast, statistical approaches can in most cases work fully automatically, and are hence much better suitable for batch processing. Second, the applied physical models and analysis methods can be explicitly checked for their approximation error. This

makes physics-based methods inherently explainable, which is an excellent property for defending a decision based on a physics-based analysis. The majority of physics-based approaches do not use advanced machine learning methods, since this would make the explainability of the results much more complicated. Third, physics-based algorithms require specific scene constellations in order to be applicable. For example, an analysis that assumes that the sun is the only light source in the scene is not applicable to indoor photographs or night time pictures. Conversely, the focus on scene content in conjunction with manual annotations by analysts makes physics-based methods by design extremely robust to the quality or processing history of an image. This is a clear benefit over statistical forensic algorithms, since their performance quickly deteriorates on images of reduced quality or complex post-processing. Some physics-based methods can even be applied to printouts, which is also not possible for statistical methods.

With these three properties, physics-based forensic algorithms can be seen as complementary to statistical approaches. Their applicability is limited to specific scenes and manual interactions, but they are inherently explainable and very robust to the processing history of an image.

Closely related to physics-based methods are behavioral cues of persons and physiological features as discussed in Chap. 11 of this book. These methods are not physics-based in the strict sense, as they do not use physical models. However, these methods share with physics-based methods the property that they operate on the scene content, and offer as such in many cases also resilience to various processing histories.

9.1.3 Outline of This Chapter

A defining property of physics-based methods is the underlying analytic models and their assumptions. The most widely used models will be introduced in Sect. 9.2. The models are divided into two parts: geometric and optical models are introduced in Sect. 9.2.1, and photometric and reflectance models are introduced in Sect. 9.2.2. Applications of these models in forensic algorithms are presented in Sect. 9.3. This part is subdivided into geometric methods, photometric methods, and combinations thereof. This chapter concludes with a discussion on the strength and weaknesses of physics-based methods and an outlook on emerging challenges in Sect. 9.4.

9.2 Physics-Based Models for Forensic Analysis

Consider a photograph of a building, for example, the Hagia Sophia in Fig. 9.1. The way how this physical object is converted to a digital image is called image formation. Three aspects of the image formation are particularly relevant for physics-based algorithms. These three aspects are illustrated below.

Fig. 9.1 Example photograph "Turkey-3019—Hagia Sophia" (full picture credits at the end of chapter)

First, landmark points like the tips of the towers are projected onto the camera. For all conventional cameras that operate with a lens, the laws of perspective projection determine the locations to which these landmarks are projected onto the sensor. Second, the dome of the building is bright where the sun is reflected into the camera, and darker at other locations. Brightness and color variations across pixels are represented with photometric models. Third, the vivid colors and, overall, the final representation of the image are obtained from the several camera-internal processing functions, which computationally perform linear and non-linear operations on the image.

The first two components of the image formation are typically used for a physics-based analysis. We introduce the foundation for the analysis of geometric properties of the scene in Sect. 9.2.1, and we introduce foundations for the analysis of reflectance properties of the scene in Sect. 9.2.2.

9.2.1 Geometry and Optics

The use of geometric laws in image analysis is a classical topic from the field of computer vision. Geometric algorithms typically examine the relation between the location of 3-D points in the scene and their 2-D projection onto the image. In computer vision, the goal is usually to infer a consistent 3-D structure from one or more 2-D images. One example is the stereo vision to calculate depth maps for an object or scene from two laterally shifted input images. In multimedia forensics, the

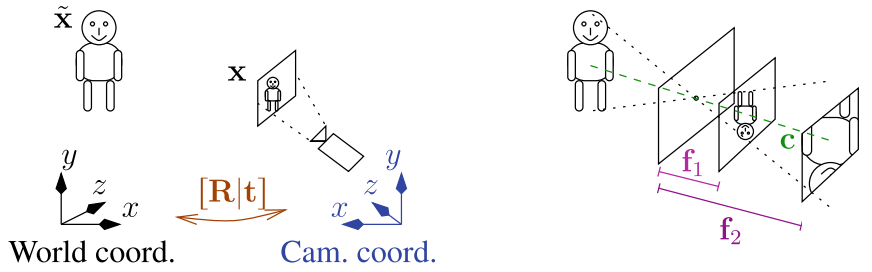

Fig. 9.2 The mapping between world- and camera coordinates is performed with the extrinsic camera parameters (left). The mapping from the 3-D scene onto the 2-D image plane is performed with the intrinsic camera parameters (right)

underlying assumption of a geometric analysis is that the geometric consistency of the scene is violated by an inserted object or by certain types of object editing.

For brevity of notation, let us denote a point in the image $\mathbf{I}$ at coordinate (y, x) as $\mathbf{x} = I(y, x)$. This point corresponds to a 3-D point in the world at the time when the image was taken. In the computer vision literature, this point is typically denoted as $\mathbf{X}$, but here it is denoted as $\tilde{\mathbf{x}}$ to avoid notational confusion with matrices. Vectors that are used in projective geometry are typically written in homogeneous coordinates. This also allows to conveniently express points at infinity. Homogeneous vectors have one extra dimension with an entry z_h, such that the usual Cartesian coordinates are obtained by dividing each other entry by z_h. Assume that we convert Cartesian coordinates to homogeneous coordinates, and choose $z_h = 1$. Then, the remaining vector elements x_i are identical to their corresponding Cartesian coordinates, since $x_i/1 = x_i$.

When taking a picture, the camera maps the 3-D point $\tilde{\mathbf{x}}$ from the 3-D world onto the 2-D image point $\mathbf{x}$. Mathematically, this is a perspective projection, which can be very generally expressed as

$$\mathbf{x} = \mathbf{K}[\mathbf{R}|\mathbf{t}]\tilde{\mathbf{x}} \ , \tag{9.1}$$

where $\mathbf{x}$ and $\tilde{\mathbf{x}}$ are written in homogeneous coordinates. This projection consists of two matrices: the matrix $[\mathbf{R}|\mathbf{t}] \in \mathbb{R}^{3\times4}$ contains the so-called extrinsic parameters. It transforms the point $\tilde{\mathbf{x}}$ from an arbitrary world coordinate system to the 3-D coordinates of the camera. The matrix $K \in \mathbb{R}^{3\times3}$ contains the so-called intrinsic parameters. It maps these 3-D coordinates onto the 2-D image plane of the camera.

Both mapping steps are illustrated in Fig. 9.2. The mapping of the world coordinate system to the camera coordinate system via the extrinsic camera parameters is shown on the left. The mapping of the 3-D scene onto the 2-D image plane is shown on the right.

Both matrices have a special form. The matrix of extrinsic parameters $[\mathbf{R}|\mathbf{t}]$ is a 3×4 matrix. It consists of a 3×3 rotation matrix in the first three columns and a 3×1 translation vector in the fourth column. The matrix of intrinsic parameters K

is an upper triangular 3×3 matrix

$$K = \begin{pmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix} , \tag{9.2}$$

where f_x and f_y are the focal length in pixels along x- and y-direction, (c_x, c_y) is the camera principal point, and s denotes pixel skewness. The camera principal point (oftentimes also called camera center) is particularly important for forensic applications, as it marks the center of projection. The focal length is also of importance, as it changes the size of the projection cone from world coordinates to image coordinates. This is illustrated in the right part of Fig. 9.2 with two example focal lengths: the smaller focal length f_1 maps the whole person onto the image plane, whereas the larger focal length f_2 only maps the torso around the center of projection (green) onto the image plane.

Forensic algorithms that build on top of these equations typically do not use the full projection model. If only a single image is available, and no further external knowledge on the scene geometry, then the world coordinate system can be aligned with the camera coordinate system. In this case, the x- and y-axes of the world coordinate system correspond to the x- and y-axes of the camera coordinate system, and the z-axis points with the camera direction into the scene. In this case, the matrix $[\mathbf{R}|\mathbf{t}]$ can be omitted. Additionally, it is a common assumption to assume that the pixels are square, the lens is sufficiently homogeneous, and that the camera skew is negligible. These assumptions simplify the projection model to only three unknown intrinsic parameters, namely

$$\mathbf{x} = \begin{pmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{pmatrix} \tilde{\mathbf{x}} . \tag{9.3}$$

These parameters are the focal length f and the center of projection (c_x, c_y), which are used in several forensic algorithms.

One interesting property of the perspective projection is the so-called vanishing points: in the projected image, lines that are parallel in the 3-D world converge to a vanishing point on the image. Vanishing points are not necessarily visible within the shown scene but can also be outside of the scene. For example, a 3-D building typically consists of parallel lines in three mutually orthogonal directions. In this special case, the three vanishing points span an orthogonal coordinate system with the camera principal point at the center.

One standard operation in projective geometry is homography. It maps points from one plane onto another. One example application is shown in Fig. 9.3. On the left, a perspectively distorted package of printer paper is shown. The paper is in A4 format, and has hence a known ratio between height and width. After annotation of the corner points (red squares in the left picture), the package can be rectified to obtain a virtual frontal view (right picture). In forensic applications, we usually use

Fig. 9.3 Example homography. Left: rectification of an object with known aspect ratio (the A4 paper) using four annotated corners (red dots). Right: rectified object after calculation of the homography from the distorted paper plane onto the world coordinate plane

a mapping from a plane in the 3-D scene onto the camera image plane or vice versa as presented in Sect. 9.3.1. For this reason, we review the homography calculation in the following paragraphs in greater detail.

All points on a plane lie on a 2-D manifold. Hence, the mapping of a 2-D plane onto another 2-D plane is actually a transform of 2-D points onto 2-D points. We denote by $\hat{\mathbf{x}}$ a point on a plane in homogeneous 2-D world coordinates, and a point in homogeneous 2-D image coordinates as $\mathbf{x}$. Then, the homography is performed by multiplication with a 3×3 projection matrix $\mathbf{H}$,

$$\mathbf{x} = \mathbf{H}\hat{\mathbf{x}} \ . \tag{9.4}$$

Here, the equality sign holds only up to the scale of the projection matrix. This scale ambiguity does not affect the equation, but it implies that the homography matrix $\mathbf{H}$ has only 8 degrees of freedom instead of 9. Hence, to estimate the homography matrix, a total of 8 constraints are required. These constraints are obtained from point correspondences. A point correspondence is a pair of two matching points, where one point lies on one plane and the second point on the other plane, and both points mark the same location on the object. The x-coordinates and the y-coordinates of a correspondence, each contribute one constraint, such that a total of four point correspondences suffices to estimate the homography matrix.

The actual estimation is performed by solving Eq. 9.4 for the elements $\mathbf{h} = (h_{11}, \ldots, h_{33})^{\mathrm{T}}$ of $\mathbf{H}$. This estimation is known as Direct Linear Transformation, and can be found in computer vision textbooks, e.g., by Hartley and Zisserman (2013, Chap. 4.1). The solution has the form

$$\mathbf{A}\mathbf{h} = \mathbf{0} \ , \tag{9.5}$$

where $\mathbf{A}$ is a $2 \cdot N \times 9$ matrix. Each of the N available point correspondences contributes two rows to $\mathbf{A}$ of the form

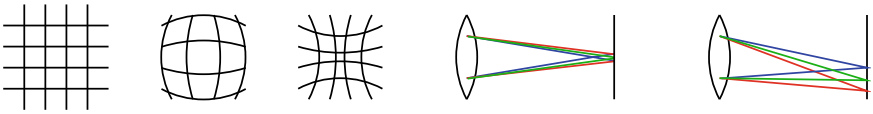

Fig. 9.4 Distortions and aberrations from the camera optics. From left to right: image of a rectangular grid, and its image under barrel distortion and pincushion distortion, axial chromatic aberration and lateral chromatic aberration

$$\mathbf{a}_{2k} = (-\hat{x}_1, -\hat{y}_1, -1, 0, 0, 0, x_2\hat{x}_1, x_2\hat{x}_2, x_2) \tag{9.6}$$

and

$$\mathbf{a}_{2k+1} = (0, 0, 0, -\hat{x}_1, -\hat{y}_1, -1, y_2\hat{x}_1, y_2\hat{x}_2, y_2) \ . \tag{9.7}$$

Equation 9.5 can be solved via singular value decomposition (SVD). After the factorization into $A = U\Sigma V^{\mathrm{T}}$, the unit singular vector that corresponds to the smallest singular value is the solution for $\mathbf{h}$.

The homography matrix implicitly contains the intrinsic and extrinsic camera parameters. More specifically, the RQ-decomposition factorizes a matrix into a product of an upper triangular matrix and an orthogonal matrix (Hartley and Zisserman 2013, Appendix 4.1). Applied to the 3×3 matrix $\mathbf{H}$, the 3×3 upper triangular matrix corresponds then to the intrinsic camera parameter matrix $\mathbf{K}$. The orthogonal matrix corresponds to a 3×3 matrix $[\mathbf{r}_1\mathbf{r}_2\mathbf{t}]$, where the missing third rotation vector of the extrinsic parameters is calculated as the cross-product $\mathbf{r}_3 = \mathbf{r}_1 \times \mathbf{r}_2$.

The equations above implicitly assume a perfect mapping from points in the scene to pixels in the image. However, in reality, the camera lens is not perfect, which can slightly alter the pixel representation. We briefly state notable deviations from a perfect mapping without going further into detail. Figure 9.4 illustrates these deviations. The three grids on the left illustrate geometric lens distortions, which change the mapping of lines from the scene onto the image. Probably, the most well-known types are barrel distortion and pincushion distortion. Both types of distortions either stretch or shrink the projection radially around a center point. Straight lines appear then either stretched away from the center (barrel distortion) or contracted toward the center (pincushion distortion). While lens distortions affect the mapping of macroscopic scene elements, another form of lens imperfection is chromatic aberration, which is illustrated on the right of Fig. 9.4. Here, the mapping of a world point between the lens and image plane is shown. However, different wavelengths (and hence colors) are not mapped onto the same location on the sensor. The first illustration shows axial chromatic aberration, where individual color channels focus at different distances. In the picture, only the green color channel converges at the image plane and is in focus; the other two color channels are slightly defocused. The second illustration shows lateral chromatic aberration, which displaces different wavelengths. In extreme cases, this effect can also be seen when zooming into a high-resolution photograph as a slight color seam at object boundaries that are far from the image center.

9.2.2 Photometry and Reflectance

The use of photometric and reflectance models in image analysis is also a classical topic from the field of computer vision. These models operate on brightness or color distributions of pixels. If a pixel records the reflectance from the surface of a solid object, its brightness and color are a function of the object material, and the relative orientation of the surface patch, the light sources, and the camera in the scene. In computer vision, classical applications are, for example, intrinsic image decomposition for factorizing an object into geometry and material color (called albedo), photometric stereo and shape from shading for reconstructing the geometry of an object from the brightness distribution of its pixels, and color constancy for obtaining a canonical color representation that is independent of the color of the illumination. In multimedia forensics, the underlying assumption is that the light source and the camera induce global consistency constraints that are violated when an object is inserted or otherwise edited.

The color and brightness of a single pixel $I(y, x)$ are oftentimes modeled as irradiance of light that is reflected from a surface patch onto the camera lens. To this end, the surface patch itself must be illuminated by one or more light sources. Each light source is assumed to emit photons with a particular energy distribution, which is the spectrum of the light source. In human vision, the visible spectrum of a light source is perceived as the color of the light. The surface patch may reflect this light as diffuse or specular reflectance, or a combination of both.

Diffuse reflectance, also called Lambertian reflectance, is by far the most commonly assumed model. The light-object interaction is illustrated in the left part of Fig. 9.5. Here, photons entering the object surface are scattered within the object, and then emitted in a random direction. Photons are much more likely to enter an object when they perpendicularly hit the surface than when they hit at a flatter angle. This is mathematically expressed as a cosine between the angle of incidence and the surface normal. The scattering within the object changes the spectrum toward the albedo of the object, such that the spectrum after leaving the object corresponds to the object color given the color of the light source. Since the photons are emitted in random directions, the perceived brightness is identical from all viewing directions.

Specular reflectance occurs when photons do not enter the object, but instead are reflected at the surface. The light-object interaction is illustrated in the right part

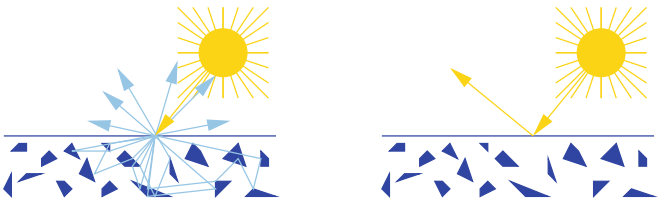

Fig. 9.5 Light-object interaction for diffuse (Lambertian) and specular reflectance

of Fig. 9.5. Specular reflectance has two additional properties. First, the spectrum of the light is barely affected by this minimal interaction with the surface. Second, specular reflectance is not scattered in all directions, but only in a very narrow angle of exitance that is opposite to the angle of incidence. This can be best observed on mirrors, which exhibit almost perfect specular reflectance.

The photometric formation of a pixel can be described by a physical model. Let $\mathbf{i} = I(y, x)$ be the usual RGB color representation at pixel $I(y, x)$. We assume that this pixel shows an opaque surface object. Then, the irradiance of that pixel into the camera is quite generally described as

$$\mathbf{i}(\boldsymbol{\theta}_\mathrm{e}, \nu) = f_\mathrm{cam}\left(\int_{\boldsymbol{\theta}_\mathrm{i}}\int_{\lambda} r(\boldsymbol{\theta}_\mathrm{i}, \boldsymbol{\theta}_\mathrm{e}, \nu, \lambda) \cdot e(\boldsymbol{\theta}_\mathrm{i}, \lambda) \cdot c_\mathrm{rgb}(\lambda)\,\mathrm{d}\,\boldsymbol{\theta}_\mathrm{i}\,\mathrm{d}\,\lambda\right) . \tag{9.8}$$

Here, $\boldsymbol{\theta}_\mathrm{i}$ and $\boldsymbol{\theta}_\mathrm{e}$ denote the 3-D angles under which a ray of light falls onto the surface and exits the surface, λ denotes the wavelength of light, ν is the normal vector of the surface point, $e(\boldsymbol{\theta}_\mathrm{i}, \lambda)$ denotes the amount of light coming from angle $\boldsymbol{\theta}_\mathrm{i}$ with wavelength λ, $r(\boldsymbol{\theta}_\mathrm{i}, \boldsymbol{\theta}_\mathrm{e}, \nu, \lambda)$ is the reflectance function indicating the fraction of light that is reflected in direction $\boldsymbol{\theta}_\mathrm{e}$ after it arrived with wavelength λ from angle $\boldsymbol{\theta}_\mathrm{i}$, and $c_\mathrm{rgb}(\lambda)$ models the camera's color sensitivity in the three RGB channels to wavelength λ. The inner expression integrates over all angles of incidence and all wavelengths to model all light that potentially leaves this surface patch in the direction of the camera. The function f_cam captures all further processing in the camera. For example, consumer cameras always apply a non-linear scaling called gamma factor to each individual color channel to make the colors appear more vivid. Theoretically, f_cam could also include additional processing that involves multiple pixels, such as demosaicking, white balancing, and lens distortion correction, although this is of minor concern for the algorithms in this chapter.

The individual terms of Eq. 9.8 are illustrated in Fig. 9.6. Rays of light are emitted by the light source and arrive at a surface with term $e(\boldsymbol{\theta}_\mathrm{i}, \lambda)$. Reflections from the surface with function $r(\boldsymbol{\theta}_\mathrm{i}, \boldsymbol{\theta}_\mathrm{e}, \nu, \lambda)$ are filtered for their colors on the camera sensor with c_rgb, and further processed in the camera with f_cam.

For algorithm development, Eq. 9.8 is in most cases too detailed to be useful. Hence, most components are typically neutralized by additional assumptions. For example, algorithms that focus on geometric arguments, such as the analysis of lighting environments, typically assume to operate on a grayscale image (or just a single-color channel), and hence remove all influences from the wavelength λ and the color sensitivity c_rgb from the model. Conversely, algorithms that focus on color typically ignore the integral over $\boldsymbol{\theta}_\mathrm{i}$ and all other influences of $\boldsymbol{\theta}_\mathrm{i}$ and $\boldsymbol{\theta}_\mathrm{e}$, and also assume a greatly simplified color sensitivity function c_rgb. Both types of algorithms oftentimes assume that the camera response function f_cam is linear. In this case, it is particularly important to invert the gamma factor as a pre-processing step when analyzing pictures from consumer cameras. The exact inversion formula depends on

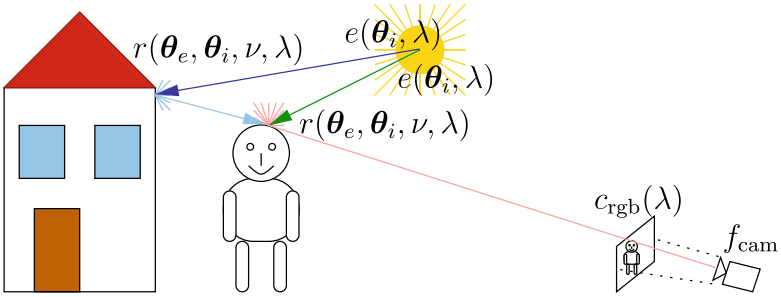

Fig. 9.6 Photometric image formation. Light sources emit light rays, which arrive as $e(\boldsymbol{\theta}_\text{i}, \lambda)$ with wavelength λ at an angle of incidence $\boldsymbol{\theta}_\text{i}$ on a surface. The surface reflectance $r(\boldsymbol{\theta}_\text{i}, \boldsymbol{\theta}_\text{e}, \nu, \lambda)$ models the reflection. Several of these rays may mix and eventually reach the camera sensor with color filters c_{rgb}. Further in-camera processing is denoted by f_{cam}

the color space. For the widely used sRGB color space, the inversion formula for a single RGB-color channel i_{sRGB} of a pixel is

$$i_{\text{linearRGB}} = \begin{cases} \frac{25 i_{\text{sRGB}}}{323} & \text{if } i_{\text{sRGB}} < 0.04045 \\ \left(\frac{200 i_{\text{sRGB}} + 11}{211}\right)^{\frac{12}{5}} & \text{otherwise} \end{cases} \tag{9.9}$$

where we assume that the intensities are in a range from 0 to 1.

The reflectance $r(\boldsymbol{\theta}_\text{i}, \boldsymbol{\theta}_\text{e}, \lambda)$ is oftentimes assumed to be purely diffuse, i.e., without any specular highlights. This reflectance model is called Lambertian reflectance. In this simple model, the angle of exitance $\boldsymbol{\theta}_\text{e}$ is ignored. The amount of light that is reflected from an object only depends on the angle of incidence $\boldsymbol{\theta}_\text{i}$ and the surface normal ν. More specifically, the amount of reflected light is the cosine between the angle of incidence and the surface normal ν of the object at that point. The full reflectance function also encodes the color of the object, written here as $s_d(\lambda)$, which yields a product of the cosine due to the ray geometry and the color,

$$r_{\text{Lambertian}}(\boldsymbol{\theta}_\text{i}, \nu, \lambda) = \cos(\boldsymbol{\theta}_\text{i}, \nu) s_d(\lambda) \ . \tag{9.10}$$

In many use cases of this equation, it is common to pull $s_d(\lambda)$ out of the equation (or to set it to unity, thereby ignoring the impact of color) and to only consider the geometric term $\cos(\boldsymbol{\theta}_\text{i}, \nu)$.

The dichromatic reflectance model is a linear combination of purely diffuse and specular reflectance, i.e.,

$$r_{\text{dichromatic}}(\boldsymbol{\theta}_\text{i}, \boldsymbol{\theta}_\text{e}, \nu, \lambda) = \cos(\boldsymbol{\theta}_\text{i}, \nu) s_d(\lambda) + w_s(\boldsymbol{\theta}_\text{i}, \nu, \boldsymbol{\theta}_\text{e}) s_s(\lambda) \ . \tag{9.11}$$

Here, the purely diffuse term is identical to the Lambertian reflectance Eq. 9.10. The specular term again decomposes into a geometric part w_s and a color part s_s. Both are to my knowledge not explicitly used in forensics, and can hence be superficially

treated. However, the geometric part w_s essentially contains the mirror equation, i.e., the angle of incidence $\boldsymbol{\theta}_\mathrm{i}$ and the angle of exitance $\boldsymbol{\theta}_\mathrm{e}$ have to be mirrored around the surface normal. The object color s_s is typically set to unity with an additional assumption, which is called the neutral interface reflectance function. This has the effect that the color of specular highlights is equal to the color of the light source when inserting the dichromatic reflectance function $r_{\mathrm{dichromatic}}$ into the full photometric model in Eq. 9.8.

9.3 Algorithms for Physics-Based Forensic Analysis

This section reviews several forensic approaches that exploit the introduced physical models. We introduce in Sect. 9.3.1 geometric methods that directly operate on the physical models from Sect. 9.2.1. We introduce in Sect. 9.3.2 photometric methods that exploit the physical models from Sect. 9.2.2. Then, we present in Sect. 9.3.3 methods that slightly relax assumptions from both geometric and photometric approaches.

9.3.1 Principal Points and Homographies

We start with a method that extracts relatively strong constraints from the scene. Consider a picture that contains written text, for example, signs with political messages during a rally. If a picture of that sign is not taken exactly frontally, but at an angle, then the letters are distorted according to the laws of perspective projection. Conotter et al. proposed a direct application of homographies to validate a proper perspective mapping of the text (Conotter et al. 2010): An attacker who aims to manipulate the text on the sign must also distort the text. However, if the attacker fits the text only in a way that is visually plausible, the laws of perspective projection are still likely to be violated.

To calculate the homography, it is necessary to obtain a minimum of four point correspondences between the image coordinates and a reference picture in world coordinates. However, if only a single picture is available, there is no outside reference for the point correspondences. Alternatively, if the text is printed with a commonly used font, the reference can be synthetically created by rendering the text in that font. From this reference and the text in the image are SIFT keypoints extracted to obtain point correspondences for homography calculation. The matrix $\mathbf{A}$ from Eq. 9.5 is composed with at least four corresponding keypoints points and solved. To avoid degenerate solutions, the corresponding points must be selected such that there are always at least two corresponding points that do not lie on a line with other corresponding points. The homography matrix $\mathbf{H}$ can then be used to perform the inverse homography, i.e., from image coordinates to world coordinates, and to calculate the root mean square error (RMSE) between the reference and the transformed image.

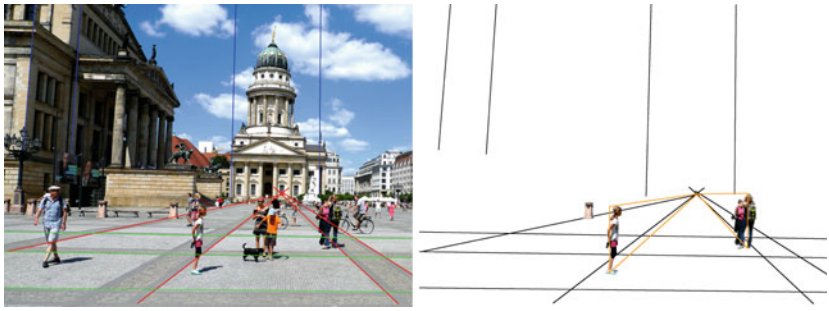

Fig. 9.7 Example scene for height measurements ("berlin square" by zoetnet; full picture credits at the end of the chapter). Lines that converge to the three vanishing points are shown in red, green, and blue. With an object or person with known height, the heights of other objects or persons can be calculated

This method benefits from its relatively strong scene constraints. If it is plausible to assume that the written text is indeed on a planar surface, and if a sufficiently similar font is available (or even the original font), then it suffices to calculate the reprojection error of a homography onto the reference. A similar idea can be used if two images show the same scene, and one of the images has been edited (Zhang et al. 2009a). In this case, the second image serves as a reference for the mapping of the first image. However, many scenes do not provide an exact reference like the exact shape of the written text or a second picture. Nevertheless, if there is other knowledge about the scene, slightly more complex variations of this idea can be used.

For example, Yao et al. show that the vanishing line of a ground plane perpendicular to the optical axis can be used to calculate the height ratio of two persons or objects at an identical distance to the camera (Yao et al. 2012). This height ratio may be sufficient when additional prior knowledge is available like the actual body height of the persons. Iuliani et al. generalize this approach to ground planes with non-zero tilt angles (Iuliani et al. 2015). This is illustrated in Fig. 9.7. The left side shows in red, green, and blue reference lines from the rectangular pattern on the floor and vertical building structures to estimate the vanishing points. The right side shows that the height of the persons on the reference plane can then be geometrically related. Note two potential pitfalls in this scene: first, when the height of persons is related, it may still be challenging to compensate for different body poses (Thakkar and Farid 2021). Second, the roadblock on the left could in principle provide the required reference height. However, it is not exactly on the same reference plane, as the shown public square is not entirely even. Hence, measurements with that block may be wrong.

Another example is to assume some prior information about the geometry of objects in the scene. Then, a simplified form of camera calibration can be performed on each object. In a second step, all objects can be checked for their agreement on the calibration parameters. If two objects disagree, it is assumed that one of these objects has been inserted from another picture with different calibration parameters.

For example, Johnson and Farid proposed to use the eyes of persons for camera calibration (Johnson and Farid 2007a). The underlying assumption is that a specific part, namely the eyes' limbi lie on a plane in 3-D space. The limbus is the boundary between the iris and the white of the eye. The limbi are assumed to be perfect circles when facing the camera directly. When viewed at an angle, the appearance of the limbi is elliptical. Johnson and Farid estimate the homography from these circles instead of isolated points. Since the homography also contains information about the camera intrinsic parameters, the authors calculate the principal point under the additional assumptions that the pixel skew is zero, and that the focal length is known from metadata or contextual knowledge of the analyst. In authentic photographs, the principal point can be assumed to be near the center of the image. If the principal point largely deviates from the image center, it can be plausibly concluded that the image is cropped (Xianzhe et al. 2013; Fanfani et al. 2020). Moreover, if there are multiple persons in the scene with largely different principal points, it can be concluded that the image is spliced from two sources. In this case, it is likely that the relative position of one person was changed from the source image to the target image, e.g., by copying that person from the left side of the image to the right side. Another useful type of additional knowledge can be structures with orthogonal lines like man-made buildings. These orthogonal lines can be used to estimate their associated vanishing points, which also provides the camera principal point (Iuliani et al. 2017).

An inventive variation of these ideas has been used by Conotter et al. (2012). They investigate ballistic motions in videos, as it occurs, e.g., for a video of a thrown basketball. Here, the required geometric constraint does not come from an object per se, but instead from the motion pattern of an object. The assumption of a ballistic motion pattern includes a linear parabolic motion without external forces except for initial acceleration and gravity. This also excludes drift from wind. Additionally, the object is assumed to be rigid and compact with a well-defined center of mass. Under these assumptions, the authors show that the consistency of the motion can be determined by inserting the analytic motion equation into the perspective projection Eq. 9.1. This model holds not only for a still camera, but also for a moving camera if the camera motion can be derived from additional static objects in the surrounding. The validation of true physical motion is additionally supported by the cue that the projection of the object size becomes smaller when the object moves away from the camera and vice versa.

Kee and Farid (2009) and Peng et al. (2017a) use the additional assumption that 3-D models are known for the heads of persons in a scene. This is a relatively strong assumption, but such a 3-D model could, for example, be calculated for people of public interest from which multiple photographs from different perspectives exist, or a 3-D head model could be captured as part of a court case. With this assumption, Kee and Farid show that approximately co-planar landmarks from such a head model can also be used to estimate a homography when the focal length can be retrieved from the metadata (Kee and Farid 2009). Peng et al. use the full 3-D set of facial landmarks and additional face contours to jointly estimate the principal point and the focal length (Peng et al. 2017a). Their main contribution is to show that spliced

images of faces can be exposed when the original faces have been captured with different focal length settings.

9.3.2 Photometric Methods

Pixel colors and brightness offer several cues to physics-based forensic analysis. In contrast to the approaches of the previous section, the photometric methods do not model perspective projections, but instead argue about local intensity distributions of objects. We distinguish two main directions of investigation, namely the distribution of the amount of light that illuminates an object from various directions and the color of light.

9.3.2.1 Lighting Environments

Johnson and Farid proposed the first method to calculate the distribution of incident light on an object or person (Johnson and Farid 2005, 2007b). We discuss this method in greater detail due to its importance in the field of physics-based methods. The method operates on a simplified form of the generalized irradiance model from Eq. 9.8. One simplifying assumption is that the camera function is linear, or that a non-linear camera response has been inverted in a pre-processing step. Another simplifying assumption is that the algorithm is only used on objects of a single color, hence all dependencies of Eq. 9.8 on wavelength λ can be ignored. Additionally, Lambertian reflectance is assumed, which removes the dependency on $\boldsymbol{\theta}_\mathrm{e}$. These assumptions lead to the irradiance model

$$\mathbf{i}(\nu) = \int_{\boldsymbol{\theta}_\mathrm{i}} \cos(\boldsymbol{\theta}_\mathrm{i}, \nu) \cdot e(\boldsymbol{\theta}_\mathrm{i}) \, \mathrm{d}\, \boldsymbol{\theta}_\mathrm{i} \tag{9.12}$$

for a single image pixel showing object surface normal ν. Here, we directly inserted Eq. 9.10 for the Lambertian reflectance model without the wavelength λ, i.e., the color term in $s_d(\lambda)$ in Eq. 9.10 is set to unity.

Johnson and Farid present two coupled ideas to estimate the distribution of incident light: First, the integral over the angles of incident rays can be summarized by only 9 parameters in the orthonormal spherical harmonics basis. Second, these nine basis coefficients can be easily regressed when a minimum of 9 surface points are available where both the intensity $\mathbf{i}$ and their associated surface normal vectors ν are known.

Spherical harmonics are a frequency representation of intensities on a sphere. In our case, they are used to model the half dome of directions from which light might fall onto a surface point of an opaque object. We denote the spherical harmonics basis functions as $h_{i,j}(x, y, z)$ where $j \leq 2i - 1$ and the parameters x, y, and z are points on a unit sphere, i.e., $\|(x, y, z)^\mathrm{T}\|_2^2 = 1$. Analogous to other frequency

transforms like the Fourier transform or the DCT, the zeroth basis function $h_{0,0}$ is the DC component that contains the overall offset of the values. Higher orders, i.e., where $i > 0$, contain increasingly higher frequencies. However, it suffices to consider only the basis functions up to second order (i.e., $i \leq 2$), to model all possible intensity distributions that can be observed under Lambertian reflectance. These orders $i = \{0, 1, 2\}$ consist of a total of 9 basis functions. For these 9 basis functions, the coefficients are estimated in the proposed algorithm.

The solution for the lighting environment requires knowledge about the surface normals from image locations where the brightness is measured. Obtaining such surface normals from only a 2-D image can be difficult. Johnson and Farid propose two options to address this challenge. First, if the 3-D structure of an object is known, a 3-D model can be created and fitted to the 2-D image. Then, the normals from the fitted 3-D model can be used. This approach has been demonstrated by Kee and Farid for faces, for which reliable 3-D model fitting methods are available (Kee and Farid 2010). Second, if no explicit knowledge about the 3-D structure of an object exists, then it is possible to estimate surface normals from occluding contours of an object. At an occluding contour, the surface normal is approximately coplanar with the image plane. Hence, the z-component is zero, and the x- and y-components can be estimated as lines that are orthogonal to the curvature of the contour. However, without the z-component, also the estimated lighting environment can only be estimated as a projection onto the 2-D image plane. On the other hand, the spherical harmonics model also becomes simpler. By setting all coefficients that contain z-components to zero, only 5 unknown coefficients remain (Johnson and Farid 2005, 2007b).

The required intensities can be directly read from the pixel grid. Johnson and Farid use the green color channel, since this color channel is usually most densely sampled by the Bayer pattern, and it has a high sensitivity to brightness differences. When 2-D normals from occluding contours are used, the intensities in the actual pixel location at the edge of an object might be inaccurate due to sampling errors and in-camera processing. Hence, in this case the intensity is extrapolated from the nearest pixels within the object along the line of the normal vector.

The normals and intensities from several object locations are the known factors in a linear system of equations

$$\mathbf{A}\mathbf{l} = \mathbf{i} \ , \tag{9.13}$$

where $\mathbf{i} \in \mathbb{R}^{N \times 1}$ are the observed intensities at N pixel locations on an object. For these N locations, matching surface normals must be available and all constraints must be satisfied. In particular, the N locations must be selected from the same surface material, and they must be directly illuminated. $\mathbf{A}$ is the matrix of the spherical harmonics basis functions. In the 3-D case, its shape is $\mathbb{R}^{N \times 9}$, and in the 2-D case it is $\mathbb{R}^{N \times 5}$. Each of the N rows in this matrix evaluates the basis functions for the surface normal of the associated pixel. The vectors $\mathbf{l}$ are the unknown coefficients of the basis functions, with dimension $\mathbb{R}^{9 \times 1}$ in the 3-D case and $\mathbb{R}^{5 \times 1}$ in the 2-D case. An additional Tikhonov regularizer dampens higher frequency spherical harmonics coefficients, which yields the objective function

$$E(\mathbf{l}) = \|\mathbf{A}\mathbf{l} - \mathbf{i}\|_2^2 + \mu\|\mathbf{R}\mathbf{l}\|_2^2 \ , \tag{9.14}$$

with the Tikhonov regularizer $\mathbf{R} = \text{diag}(1\ 2\ 2\ 2\ 3\ 3\ 3\ 3\ 3)$ for the 3-D case, and $\mathbf{R} = \text{diag}(1\ 2\ 2\ 3\ 3)$ for the 2-D case. A minimum is found after differentiation with respect to $\mathbf{l}$ and solving for $\mathbf{l}$, i.e.,

$$\mathbf{l} = (\mathbf{A}^{\mathrm{T}}\mathbf{A} + \mu\mathbf{R}^{\mathrm{T}}\mathbf{R})^{-1}\mathbf{A}^{\mathrm{T}}\mathrm{i} \ . \tag{9.15}$$

The lighting environments of two objects can be directly compared via correlation. A simple approach is to render two spheres from the coefficients of each object, and to correlate the rendered intensities in the 3-D case, or just the intensities along the boundary of the sphere in the 2-D case. Alternatively, the coefficients can also be directly compared after a suitable transform (Johnson and Farid 2005, 2007b).

This first work on the analysis of lighting environments inspired many follow-up works to relax the relatively restrictive assumptions of the method. Fan et al. (2012) explored shape-from-shading as an intermediate step to estimate 3-D surface normals when an explicit 3-D object model is not available. Riess et al. added a reflectance term to the 2-D variant of the algorithm, such that also contours from different materials can be used (Riess et al. 2017). Carvalho et al. investigate human annotations of 3-D surface normals on objects other than faces (Carvalho et al. 2015). Peng et al. automate the 3-D variant by Kee and Farid (2010) via automated landmark detection (Peng et al. 2016). Seuffert et al. show that high-quality 3-D lighting estimation requires a good-fitting geometric model (Seuffert et al. 2018). Peng et al. further include a texture term for 3-D lighting estimation on faces (Peng et al. 2015, 2017b).

Some methods also provide alternatives to the presented models. The assumption of an orthographic projection has also been relaxed, and will be separately discussed in Sect. 9.3.3. Matern et al. integrate over the 2-D image gradients, which provides a slightly less accurate, but very robust 2-D lighting estimation (Matern et al. 2020). Huang et al. propose to calculate 3-D lighting environments from general objects using surface normals from shape-from shading algorithms (Huang and Smith 2011). However, this classic computer vision approach requires relatively simple objects such as umbrellas to be robustly applicable. Zhou et al. train a neural network to learn the lighting estimation from faces in a fully data-driven manner (Zhou et al. 2018).

An example 2-D lighting estimation is shown on top of Fig. 9.8. The person is wearing a T-shirt, which makes it difficult to find surface normals that are of the same material and at the same time point in a representative number of directions. Contours for the T-shirt and the skin are annotated in green and red, together with their calculated normals. The x-axes of the scatterplots contain the possible angles of the surface normals between $-\pi$ and $+\pi$. The y-axes contain the image intensities. The left scatterplot naively combines the intensities of the black T-shirt (green) and the light skin (red). Naively fitting the spherical harmonics to the distribution of both materials leads to a failure case: the light skin dominates the estimation, such that the dominant light source (maximum of the red line) appears to be located below the person. On the other hand, using only the skin pixels or only the T-shirt pixels leads to

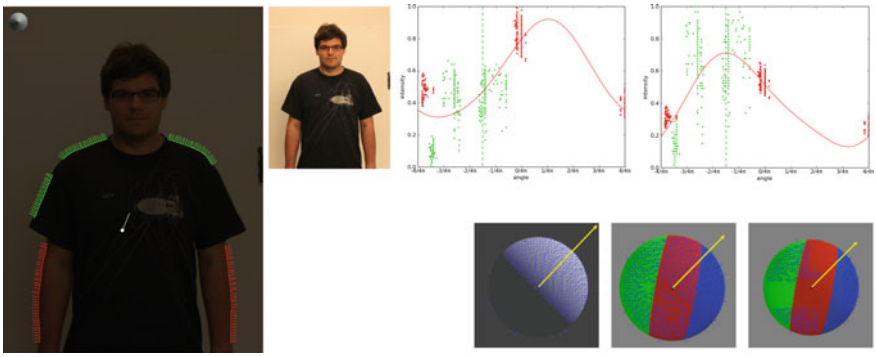

Fig. 9.8 2-D lighting estimation on objects with multiple materials. Top: Example 2-D lighting estimation according to Johnson and Farid (2005, 2007b). From left to right: contour annotations and estimated normals for two materials (green and red lines) on a person. The scatterplots show the angle of the normal along the x-axis with the associated pixel intensities along the y-axis. The left scatterplot mixes both materials, which leads to a wrong solution. The right scatterplot performs an additional reflectance normalization (Riess et al. 2017), such that both materials can be used simultaneously. Bottom: 2-D lighting environments can also be estimated from image gradients, which are overall less accurate, but considerably more robust to material transitions and wrong or incomplete segmentations

a very narrow range of surface normals, which also makes the estimation unreliable. The right plot shows the same distribution, but with the reflectance normalization by Riess et al. (2017), which correctly estimates the location of the dominant light source above the person. The three spheres on the bottom illustrate the estimation of the light source from image gradients according to Matern et al. (2020). This method only assumes that objects are mostly convex, and that the majority of local neighborhoods for gradient computation consists of the same material. With these modest assumptions, the method can oftentimes still operate on objects with large albedo differences, and on object masks with major errors in the segmentation.

9.3.2.2 Color of Illumination

The spectral distribution of light determines the color formation in an image. For example, the spectrum of sunlight is impacted by the light path through the atmosphere, such that sunlight in the morning and in the evening is more reddish than at noon. As another example, camera flashlight oftentimes exhibits a strong blue component. Digital cameras typically normalize the colors in an image with a manufacturer-specific white-balancing function that oftentimes uses relatively simple heuristics. Post-processing software such as Adobe Lightroom provides more sophisticated functions for color adjustment.

These many influencing factors on the colors of an image make their forensic analysis interesting. It is a reasonable assumption that spliced images exhibit differ-

ences in the color distribution if the spliced image components had different lighting conditions upon acquisition or if they had undergone different post-processing.

Forensics analysis methods assume that the camera white-balancing is a global image transformation that does not introduce local color inconsistencies in an original image. Hence, local inconsistencies in the color formation are attributed to potential splicing manipulations.

There are several possibilities to locally analyze the consistency of the lighting color. One key requirement is the need to estimate the illuminant color locally on individual objects to expose inconsistencies. Particularly well-suited for local illumination color estimation is the dichromatic reflectance model from Eq. 9.11 in combination with the neutral interface reflectance function. This model includes diffuse and specular reflectance, with the additional assumption that the specular portion exhibits the color of the light source.

To our knowledge, the earliest method that analyzes reflectance has been presented by Gholap and Bora (2008). This work proposes to calculate the intersection of so-called dichromatic lines of multiple objects to expose splicing, using a classical result by Tominaga and Wandell (1989): on a monochromatic object, the distribution of diffuse and specular pixels forms a 2-D plane in the 3-D RGB space. This plane becomes a line when projecting it onto the 2-D r-g chromaticity space, where $r = i_R/(i_R + i_G + i_B)$ and $g = i_G/(i_R + i_G + i_B)$ denote the red and green color channels, normalized via division by the sum of all color channels of a pixel. If the scene is illuminated by a single, global illuminant, the dichromatic lines of differently colored objects all intersect at one point. Hence, Gholap and Bora propose to check for three or more scene objects whether their dichromatic lines intersect in a single point. This approach has the advantage that it is very simple. However, to our knowledge it barely provides possibilities to validate the model assumptions, i.e., ways to check whether the assumption of dichromatic reflectance and a single global illuminant holds for the objects under investigation.

Riess and Angelopoulou (2010) propose the inverse-intensity chromaticity (IIC) space by Tan et al. (2004) to directly estimate the color of the illuminant. The IIC space is simple to calculate, but provides very convenient properties for forensic analysis. It also operates on surfaces of dichromatic reflectance, and requires only a few pixels of identical material that exhibit a mixture of specular and diffuse reflectance. For each color channel, the IIC space is calculated as a 2-D chart for related pixels. Each pixel is transformed to a tuple

$$i_{R,G,B} \rightarrow \left(\frac{1}{i_{R,G,B}}, \frac{i_{R,G,B}}{i_R + i_G + i_B} \right) , \tag{9.16}$$

where again i_R, i_G, and i_B denote the red, green, and blue color channels. On pixels with different portions of specular and diffuse reflectance, the distribution forms a triangle or a line that indicates the chromaticity of the light source at the y-axis intercept. This is illustrated in Fig. 9.9. For persons, the nose region is oftentimes a well-suited location to observe partially specular pixels. Close-ups of the manual annotations are shown together with the distributions in IIC space. In this original

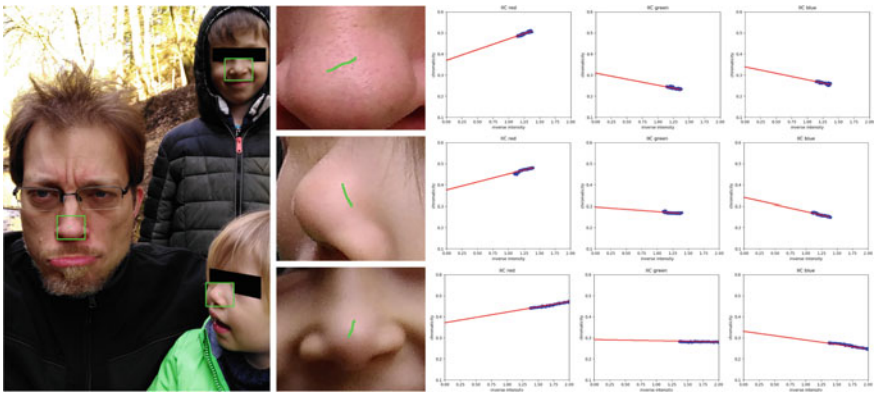

Fig. 9.9 Example application of illuminant color estimation via IIC color charts. For each person in the scene, a few pixels are selected that are of the same material, and that exhibit a mixture of specular and diffuse reflectance. The associated IIC diagrams are shown on the right. The pixel distributions (blue) point toward the estimated color of the light source at the y-axis intercept (red lines)

image, the pixel distributions for the red, green, and blue IIC diagrams point to almost identical illuminant chromaticities. In real scenes, the illuminant colors are oftentimes achromatic. However, the IIC diagrams are very sensitive to post-processing. Thus, if an image is spliced from sources with different post-processing, major differences in the IIC diagrams may be observed.

A convenient property of IIC diagrams is their interpretability. Pixel distributions that do not match the assumptions do not form a straight line. For example, purely diffuse pixels tend to cluster in circular shapes, and pixels of different materials form several clusters, which indicates to an analyst the need to revise the segmentation.

One disadvantage of the proposed method is the need for manual annotations, since the automated detection of specularities is a severely underconstrained problem, and hence not reliable. To mitigate this issue, Riess and Angelopoulou propose to estimate the color of the illuminant on small, automatically segmented image patches of approximately uniform chromaticity. They introduce the notion of "illuminant map" for a picture where each image patch is colored with the estimated illuminant chromaticity. An analyst can then consider only those regions that match the assumptions of the approach, i.e., that exhibit partially specular and partially diffuse reflectance on dichromatic materials. These regions can be considered reliable, and their estimated illuminant colors can be compared to expose spliced images. However, for practical use, we consider illuminant maps oftentimes inferior to a careful, fully manual analysis.

Another analytic approach based on the dichromatic reflectance model has been proposed by Francis et al. (2014). It is conceptually similar, but separates specular and diffuse pixels directly in RGB space. Several works use machine learning to automate the processing and to increase the robustness of illuminant descriptors. For example, Carvalho et al. use illuminant maps that are calculated from the IIC space, and also from the statistical gray edge illuminant estimator, to train a classifier for the authenticity assessment of human faces (de Carvalho et al. 2013). By constraining the application of the illuminant descriptors only to faces, the variety of possible surface materials is greatly restricted, which can increase the robustness of the estimator. However, this approach also performs further processing on the illuminant maps, which slightly leaves the ground of purely physics-based methods and enters the domain of statistical feature engineering. Hadwiger et al. investigate a machine learning approach to learn the color formation of digital cameras (Hadwiger et al. 2019). They use images with Macbeth color charts to learn the relationship between ground truth colors and their representations of different color camera pipelines. In different lines of work on image colors, Guo et al. investigate fake colorized image detection via hue and saturation statistics (Guo et al. 2018). Liu et al. investigate methods to assess the ambient illumination in shadow regions for their consistency by estimating the shadow matte (Liu et al. 2011).

9.3.3 Point Light Sources and Line Constraints in the Projective Space

Outdoor scenes that are only illuminated by the sun provide a special set of constraints. The sun can be approximated as a point light source, i.e., a source where all light is emitted from a single point in 3-D space. Also, indoor scenes may in special cases contain light sources that can be approximated as a single point, like a single candle that illuminates the scene (Stork and Johnson 2006).

Such a single-point light source makes the modeling of shadows particularly simple: the line that connects the tip of an object with the tip of its cast shadow also intersects the light source. Multiple such lines can be used to constrain the position of the light source. Conversely, shadows that have been artificially inserted may violate these laws of projection.

Several works make use of this relationship. Zhang et al. (2009b) and Wu et al. (2012) show that cast shadows of an object can be measured and validated against the length relationships of the persons or objects that cast the shadow. One assumption of these works is that the shadow is cast on a ground plane, such that the measurements are well comparable. Conversely, Stork and Johnson use cast shadows to verify that a scene is illuminated by a specific light source, by connecting occluding contours that likely stem from that light source (Stork and Johnson 2006).

However, these approaches require a clear correspondence between the location on the object that casts the shadow and the location where the shadow is cast to. This is not an issue for poles and other thin cylindrical objects, as the tip of the object can be easily identified. However, it is oftentimes difficult to find such correspondences on objects of more complex shapes.

O'Brien and Farid hence propose a geometric approach with much more relaxed requirements to mark specific locations for object-shadow correspondences (O'Brien and Farid 2012). The method is applicable not only to cast shadows, but also to a mixture of cast shadows and attached shadows as they occur on smooth surfaces. The idea is that if the 2-D image plane would be infinitely large, there would be a location onto which the light source is projected, even if it is outside of the actual image. Object-shadow correspondences can constrain the projected location of the light source. The more accurate an object-shadow correspondence can be identified, the stronger is the constraint. However, it is also possible to incorporate very weak constraints like attached shadows. If the image has been edited, and, for example, an object has been inserted with incorrect illumination, then the constraints from that inserted object are likely to violate the remaining constraints in the image.

The constraints are formed from half-planes in the 2-D image plane. An attached shadow separates the 3-D space into two parts, namely one side that is illuminated and one side that is in shadows. The projection of the illuminated side onto the 2-D image plane corresponds to one half-plane with a boundary normal that corresponds to the attached shadow edge. An object-shadow correspondence with some uncertainty about the exact location at the object and the shadow separates the 3-D space into a 3-D wedge of possible light source locations. Mapping this wedge onto the 2-D image plane leads to a 2-D wedge. Such a wedge can be constructed from two opposing half-planes.

An application of this approach is illustrated in Fig. 9.10. Here, the pen and the spoon exhibit two different difficulties for analyzing the shadows, which is shown in the close-ups on the top right: the shadow of the pen is very unsharp, such that it is difficult to locate its tip. Conversely, the shadow of the spoon is very sharp, but the round shape of the spoon makes it difficult to determine from which location on the spoon surface the shadow tip originates. However, both uncertainties can be modeled as half-plane constraints shown at the bottom. These constraints have a common intersection in the image plane outside of the image, shown in pink. Hence, the shadows of both objects are consistent. Further objects could now be included in the analysis, and their half-plane constraints have to analogously intersect the pink area.

In follow-up work, this approach has been extended by Kee et al. to also include constraints from object shading (Kee et al. 2013). This approach is conceptually highly similar to the estimation of lighting environments as discussed in Sect. 9.3.2.1. However, instead of assuming an orthographic projection, the lighting environments are formulated here within the framework of perspective projection to smoothly integrate with the shadow constraints.

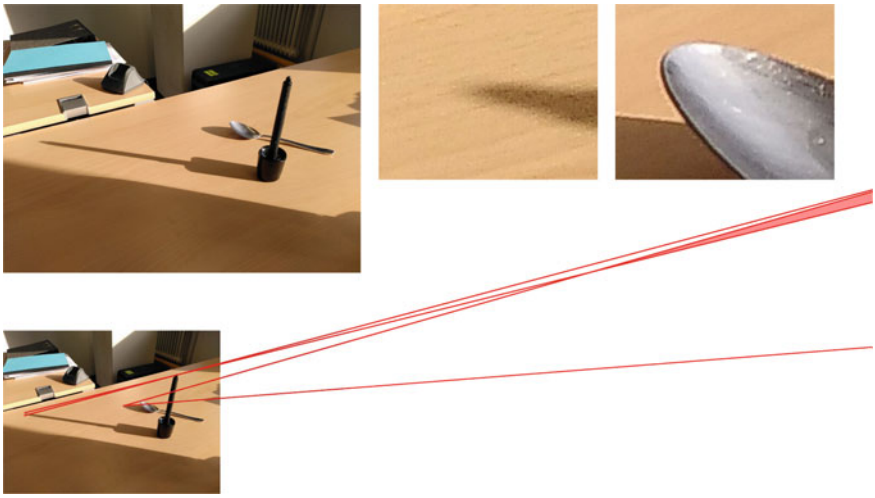

Fig. 9.10 Example scene for forensics from cast shadows. The close-ups show the uncertain areas of shadow formation. Left: the shadow of the tip of the pen is very unsharp. Right: the round shape of the spoon makes it difficult to exactly localize the location that casts the tip of the shadow. Perspective constraints according to O'Brien and Farid allow to nevertheless use these uncertainty regions for forensic analysis (O'Brien and Farid 2012): the areas of the line constraints overlap in the pink region of the image plane (although outside of the actual image), indicating that the shadows of both objects are consistent

9.4 Discussion and Outlook

Physics-based methods for forensic analysis are based on simplified physical models to validate the authenticity of a scene. Journalistic verification uses physics-based methods mostly to answer questions about the time and place of an acquisition, the academic literature mostly focuses on the detection of inconsistencies within an image or video.

Conceptually, physics-based methods are quite different from statistical approaches. Each physics-based approach requires certain scene elements to perform an analysis, while statistical methods can operate on almost arbitrary scenes. Also, most physics-based methods require the manual interaction of an analyst to provide "world knowledge", for example, to annotate occluding contours, or to select partially specular pixels. On the other hand, physics-based methods are mostly independent of the image or video quality, which makes them particularly attractive for analyzing low-quality content or even analog content. Also, physics-based methods are inherently explainable by verification of their underlying models. This would in principle also make it well possible to perform a rigorous analysis of the impact of estimation errors from various error sources, which is much more difficult for statistical approaches. Surprisingly, such robustness investigations have until now only been

performed to a limited extent, e.g., for the estimation of vanishing points (Iuliani et al. 2017), perspective distortions (Peng et al. 2017a), or height measurements (Thakkar and Farid 2021).

The future of physics-based methods is challenged by the technical progress in two directions: First, the advent of computational images in modern smartphones. Second, physically plausible computer-generated scene elements from modern methods in computer graphics and computer vision.

Computational images compensate for the limited camera optics in smartphones, which are due to space constraints not competitive with high-quality cameras. Hence, when a modern smartphone captures "an image", it actually captures a short video, and calculates a high-quality single image from that frame sequence. However, if the image itself is not the result of a physical image formation process, the validity of physics-based models for forensic analysis is inherently questioned. It is currently an open question to which extent these computations affect the presented physics-based algorithms.

Recent computer-generated scene elements are created with an increasing contribution of learned physics-based models for realistically looking virtual reality or augmented reality (VR/AR) applications. These approaches are a direct competition to physics-based forensic algorithms, as they use very similar models to minimize their representation error. This emphasizes the need for multiple complementary forensic tools to expose manipulations from cues that are not relevant for the specific tasks of such VR/AR applications, and are hence not considered in their optimization.

9.5 Picture Credits

- Figure 9.1 is published by Dennis Jarvis (archer10, https://www.flickr.com/photos/archer10/ with an Attribution-ShareAlike 2.0 Generic (CC BY-SA 2.0) License. Full text to the license is available at https://creativecommons.org/licenses/by-sa/2.0/; link to the original picture is https://www.flickr.com/photos/archer10/2216460729/. The original image is downsampled for reproduction.
- Figure 9.7 is published by zoetnet, https://flickr.com/photos/zoetnet/ with an Attributed 2.0 Generic (CC BY 2.0) License. Full text to the license is available at https://creativecommons.org/licenses/by/2.0/; link to the original picture is https://flickr.com/photos/zoetnet/9527389096/. The original image is downsampled for reproduction, and annotations of perspective lines are added.

References

Carvalho T, Farid H, Kee E (2015) Exposing photo manipulation from user-guided 3d lighting analysis. In: Alattar AM, Memon ND, Heitzenrater C (eds) Media watermarking, security, and forensics 2015, San Francisco, CA, USA, February 9–11, 2015, Proceedings, vol 9409 of SPIE Proceedings. SPIE, p 940902

Conotter V, O'Brien James F, Farid H (2012) Exposing digital forgeries in ballistic motion. IEEE Trans Inf Forensics Secur 7(1):283–296

Conotter V, Boato G, Farid H (2010) Detecting photo manipulation on signs and billboards. In: Proceedings of the international conference on image processing, ICIP 2010, September 26–29, Hong Kong, China. IEEE, pp 1741–1744

de Carvalho TJ, Riess C, Angelopoulou E, Pedrini H, de Rezende Rocha A (2013) Exposing digital image forgeries by illumination color classification. IEEE Trans Inf Forensics Secur 8(7):1182–1194

Fanfani M, Iuliani M, Bellavia F, Colombo C, Piva A (2020) A vision-based fully automated approach to robust image cropping detection. Signal Proc Image Commun 80

Fan W, Wang K, Cayre F, Xiong Z (2012) 3d lighting-based image forgery detection using shape-from-shading. In: Proceedings of the 20th European signal processing conference, EUSIPCO 2012, Bucharest, Romania, August 27–31, 2012. IEEE, pp 1777–1781

Francis K, Gholap S, Bora PK (2014) Illuminant colour based image forensics using mismatch in human skin highlights. In: Twentieth national conference on communications, NCC 2014, Kanpur, India, February 28–March 2, 2014. IEEE, pp 1–6

Gholap S, Bora PK (2008) Illuminant colour based image forensics. In: IEEE region 10 conference, TENCON 2008, Hyderabad, India, November 19–21 2008

Guo Y, Cao X, Zhang W, Wang R (2018) Fake colorized image detection. IEEE Trans Inf Forensics Secur 13(8):1932–1944

Hadwiger B, Baracchi D, Piva A, Riess C (2019) Towards learned color representations for image splicing detection. In: IEEE international conference on acoustics, speech and signal processing, ICASSP 2019, Brighton, United Kingdom, May 12–17, 2019. IEEE, pp 8281–8285

Hartley R, Zisserman A (2004) Multiple view geometry in computer vision. Cambridge University Press, Cambridge

Huang R, Smith WAP (2011) Shape-from-shading under complex natural illumination. In: Macq B, Schelkens P (eds) 18th IEEE international conference on image processing, ICIP 2011, Brussels, Belgium, September 11–14, 2011. IEEE, pp 13–16

Iuliani M, Fanfani M, Colombo C, Piva A (2017) Reliability assessment of principal point estimates for forensic applications. J Vis Commun Image Represent 42:65–77

Iuliani M, Fabbri G, Piva A (2015) Image splicing detection based on general perspective constraints. In: 2015 IEEE international workshop on information forensics and security, WIFS 2015, Roma, Italy, November 16–19, 2015. IEEE, pp 1–6

Johnson MK, Farid H (2006) Exposing digital forgeries by detecting inconsistencies in lighting. In: Eskicioglu AM, Fridrich JJ, Dittmann J (eds) Proceedings of the 7th workshop on Multimedia & Security, MM&Sec 2005, New York, NY, USA, August 1–2, 2005. ACM, pp 1–10

Johnson MK, Farid H (2007a) Detecting photographic composites of people. In: Shi YQ, Kim H-J, Katzenbeisser S (eds) Digital watermarking, 6th international workshop, IWDW 2007, Guangzhou, China, December 3–5, 2007, Proceedings, vol 5041 of Lecture notes in computer science. Springer, pp 19–33

Johnson MK, Farid H (2007b) Exposing digital forgeries in complex lighting environments. IEEE Trans Inf Forensics Secur 2(3–1):450–461

Kee E, Farid H (2009) Detecting photographic composites of famous people. Technical Report Computer Science Technical Report TR2009-656, Department of Computer Science, Dartmouth College

Kee E, Farid H (2010) Exposing digital forgeries from 3-d lighting environments. In: 2010 IEEE international workshop on information forensics and security, WIFS 2010, Seattle, WA, USA, December 12–15, 2010. IEEE, pp 1–6

Kee E, O'Brien JF, Farid H (2013) Exposing photo manipulation with inconsistent shadows. ACM Trans Graph 32(3):28:1–28:12

Liu Q, Cao X, Deng C, Guo X (2011) Identifying image composites through shadow matte consistency. IEEE Trans Inf Forensics Secur 6(3–2):1111–1122

Matern F, Riess C, Stamminger M (2020) Gradient-based illumination description for image forgery detection. IEEE Trans Inf Forensics Secur 15:1303–1317

O'Brien JF, Farid H (2012) Exposing photo manipulation with inconsistent reflections. ACM Trans Graph 31(1):4:1–4:11

Peng B, Wang W, Dong J, Tan T (2015) Improved 3d lighting environment estimation for image forgery detection. In: 2015 IEEE international workshop on information forensics and security, WIFS 2015, Roma, Italy, November 16–19, 2015. IEEE, pp 1–6

Peng B, Wang W, Dong J, Tan T (2016) Automatic detection of 3d lighting inconsistencies via a facial landmark based morphable model. In: 2016 IEEE international conference on image processing, ICIP 2016, Phoenix, AZ, USA, September 25–28, 2016. IEEE, pp 3932–3936

Peng B, Wang W, Dong J, Tan T (2017a) Position determines perspective: investigating perspective distortion for image forensics of faces. In: 2017 IEEE conference on computer vision and pattern recognition workshops, CVPR workshops 2017, Honolulu, HI, USA, July 21–26, 2017. IEEE Computer Society, pp 1813–1821

Peng B, Wang W, Dong J, Tan T (2017b) Optimized 3d lighting environment estimation for image forgery detection. IEEE Trans Inf Forensics Secur 12(2):479–494

Riess C, Angelopoulou E (2010) Scene illumination as an indicator of image manipulation. In: Böhme R, Fong PWL, Safavi-Naini R (eds) Information hiding - 12th international conference, IH 2010, Calgary, AB, Canada, June 28–30, 2010, Revised Selected Papers, vol 6387 of Lecture notes in computer science. Springer, pp 66–80

Riess C, Unberath M, Naderi F, Pfaller S, Stamminger M, Angelopoulou E (2017) Handling multiple materials for exposure of digital forgeries using 2-d lighting environments. Multim Tools Appl 76(4):4747–4764

Seuffert J, Stamminger M, Riess C (2018) Towards forensic exploitation of 3-d lighting environments in practice. In: Langweg H, Meier M, Witt BC, Reinhardt D (eds) Sicherheit 2018, Beiträge der 9. Jahrestagung des Fachbereichs Sicherheit der Gesellschaft für Informatik e.V. (GI), 25.–27.4.2018, Konstanz, volume P-281 of LNI. Gesellschaft für Informatik e.V., pp 159–169

Stork DG, Johnson MK (2006) Estimating the location of illuminants in realist master paintings computer image analysis addresses a debate in art history of the baroque. In: 18th international conference on pattern recognition (ICPR 2006), 20–24 August 2006, Hong Kong, China. IEEE Computer Society, pp 255–258

Tan RT, Nishino K, Ikeuchi K (2004) Color constancy through inverse-intensity chromaticity space. J Opt Soc Amer A 21(3):321–334

Thakkar N, Farid H (2021) On the feasibility of 3D model-based forensic height and weight estimation. In: Workshop on media forensics (in conjunction with CVPR)

Tominaga S, Wandell Brian A (1989) Standard surface-reflectance model and illuminant estimation. J Opt Soc Am A 6(4):576–584

Wu L, Cao X, Zhang W, Wang Y (2012) Detecting image forgeries using metrology. Mach Vis Appl 23(2):363–373

Xianzhe M, Ru SN, Yan Li Y (2013) Detecting photographic cropping based on vanishing points. Chinese J Electron 22(2):369–372

Yao H, Wang S, Zhao Y, Zhang X (2012) Detecting image forgery using perspective constraints. IEEE Signal Proc Lett 19(3):123–126

Zhang W, Cao X, Feng Z, Zhang J, Wang P (2009a) Detecting photographic composites using two-view geometrical constraints. In: Proceedings of the 2009 IEEE international conference on multimedia and Expo, ICME 2009, June 28–July 2, 2009, New York City, NY, USA. IEEE, pp 1078–1081

Zhang W, Cao X, Zhang J, Zhu J, Wang P (2009b) Detecting photographic composites using shadows. In: Proceedings of the 2009 IEEE international conference on multimedia and Expo, ICME 2009, June 28–July 2, 2009, New York City, NY, USA. IEEE, pp 1042–1045

Zhou H, Sun J, Yacoob Y, Jacobs DW (2018) Label denoising adversarial network (LDAN) for inverse lighting of faces. In: 2018 IEEE conference on computer vision and pattern recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018. IEEE Computer Society, pp 6238–6247