Notes on Digital Image Forensics and Counter-Forensics*

Matthias Kirchner

September 2011 / October 2012

* This text is taken from the first part of the author's dissertation, entitled *"Forensic Analysis of Resampled Digital Signals"*, submitted under supervision of Prof. Dr. rer. nat. Andreas Pfitzmann († 23.09.2010) and Prof. Dr. rer. nat. Hermann Härtig to the Department of Computer Science, TU Dresden, Germany, in early September 2011. Hermann Härtig (TU Dresden), Rainer Böhme (WWU Münster), and Hany Farid (Dartmouth College) were members of the doctoral committee. The thesis was supported by a scholarship from Deutsche Telekom Stiftung (Bonn, Germany). Parts of the third chapter were co-authored by Rainer Böhme and have been published as a book chapter *"Counter-Forensics: Attacking Image Forensics"* in the book *"Digital Image Forensics. There is More to a Picture than Meets the Eye"*, edited by Husrev T. Sencar and Nasir Nemon (Springer, 2012) [19].

Contents

1 Introduction

Since Kodak engineer Steven Sasson took the first digital image back in 1975, the world has witnessed a technological revolution that has fundamentally changed the way how images are created, consumed and perceived. Successors of Sasson's prototype—a colossus of 3.6 kilogram, which operated at a resolution of 0.01 megapixels [150]—have meanwhile matured to become omnipresent companions in our everyday life. Nowadays, easy-to-use digital imaging devices and inexpensive storage space make the acquisition of high-quality digital images a natural form of human perception of and interaction with the real world.

At the same time, the very nature of digital data puts into question many of the positive aspects that we usually associate with digital images. Digital data can be manipulated easily. Powerful editing software allows even relatively unexperienced users to conveniently process digital images in innumerable ways. While image manipulations are in general not a phenomenon exclusive to the digital world [20], critics have expressed concerns that it has never been so easy to alter content and meaning of a picture—often in such a perfection that it is impossible to visually distinguish the forgery from authentic photographs. The constantly growing number of uncovered digital image manipulations [58] signifies that the overwhelming success of digital imaging harms the trustworthiness of pictures, particularly in situations where society bases important decisions on them: in court (where photographs act as pieces of evidence), in science (where photographs provide empirical proofs), and at the ballot box (where press photographs shape public opinion).

This discrepancy between the ever-increasing relevance of digital images on the one hand and doubts regarding their susceptibility to manipulations on the other hand more than ever calls for approaches to systematically assess the trustworthiness of images. It also points to yet another ambivalence surrounding digital imagery. Not only can digital images be easily manipulated, but also give they rise to powerful computational analysis methods, which help to overcome limitations of human perception and cognition in detecting such manipulations. While the human visual system excels in grasping semantic details of a depicted scene, computational methods are typically superior at unveiling even subtlest traces at the 'syntactical' level, i. e., by examining the digital image data itself or digital data that is stored with the image [59, 26].

Farid [54] and Lukáš [151], around the turn of the millennium, first reported that post-processing leaves characteristic traces in the resulting image.[1] Around the same time, Heerich [97] and Kurosawa et al. [137] made similar endeavors to investigate inherent traces of the device that acquired the digital image. These seminal works have set the stage for the field of *digital image forensics*. Since then they have stimulated scholars from different research communities, such as multimedia security, computer forensics, imaging, and signal processing, to develop algorithms that assist the detection of manipulations and allow to infer the acquisition device of arbitrary digital images. The promise of digital image forensics is

1 Farid's [54] work deals with audio signals but was later extended to digital images by Ng et al. [176, 179].

to restore some of the lost trustworthiness of digital images. Coverage in flagship scientific journals [163], significant interest of the media as well as first practical applications in law enforcement all indicate the high relevance of digital image forensics in a digital society. Overall, we can expect this emerging discipline to play a more and more central role in the practical assessment of image trustworthiness in a broad range of real life situations.

Notation

Throughout this text, random variables defined over the domain $\mathscr{X}$ are denoted by 'plain' calligraphic letters, $\mathcal{X}$. Sets are printed in 'ornate' calligraphic letters, $\mathscr{X}$. Some special sets may also appear in a double-lined notation, e. g., $\mathbb{R}$ and $\mathbb{N}$ are the sets of all real and natural numbers, respectively. Scalars and realizations of random variables are denoted by roman lowercase or uppercase symbols, x or X. Bold symbols are reserved for vectors and matrices, respectively. A vector $\boldsymbol{x} \in \mathscr{X}^N$ holds N elements $x_i \in \mathscr{X}$, i. e., $\boldsymbol{x} = (x_1, \ldots, x_N)$. Likewise, an $M \times N$ matrix $\boldsymbol{X}$ consists of MN elements $X_{i,j} \in \mathscr{X}$ with indices $(i, j) \in \{1, \ldots, M\} \times \{1, \ldots, N\}$. A vector of random variables is denoted as $\boldsymbol{\mathcal{X}} = (\mathcal{X}_1, \ldots, \mathcal{X}_N)$. When necessary we make the dimension of a matrix explicit by writing $\boldsymbol{X}^{(M \times N)}$. More general, the superscript notation $^{(\cdot)}$ is used to make arbitrary characteristics of a vector or matrix explicit, depending on the context. Symbols M and N are reserved for dimensions of vectors and matrices, and similarly, integers K and L may denote upper limits of some sequence of indices. Single subscripts to boldface symbols (e. g., $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ or $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$) distinguish different realizations of random vectors or matrices. Functions are printed in monospaced sans serif, for example, $\mathtt{E}(x)$ or $\mathtt{process}(\boldsymbol{x})$. If not stated otherwise, we assume grayscale images $\boldsymbol{x} \in \mathbb{X}$ with bit depth ℓ, i. e, a digital image with $M \times N$ pixels is written as $\boldsymbol{x} = (x_1, \ldots, x_{MN})$, $x_i \in \mathscr{X} = [0, 2^\ell - 1]$, whereas $\mathbb{X}$ is the set of all digital images.

2 Principles of Digital Image Forensics

Digital image forensics aims at restoring some of the lost trustworthiness of digital images and revolves around the following two fundamental question:

- *Where is the image coming from?*
- *(How) Has the image been processed after acquisition?*

To answer these and related questions, digital image forensics exploits the fact that characteristics of digital images depend not only on the depicted scene but also on the particular way the image was acquired and processed. This allows to infer device(s) and post-processing steps involved in the image generation process and thereby to judge about the trustworthiness of digital images.

2.1 Forensic Analysis of Digital Images

To familiarize with the principal ideas of digital image forensics it is necessary to have at least a rough working definition of the key terms involved. Hence, we open this chapter with a short account of how different *image generation processes* (Section 2.1.1) may affect the *trustworthiness* (Section 2.1.2) of digital images. We then detail our notion of *digital image forensics* (Section 2.1.3), which is based on the examination of *image characteristics and identifying traces* (Section 2.1.4) of the respective image generation process, and finally discuss the role of *counter-forensics* (Section 2.1.5) in this setting.

2.1.1 Formation of Digital Images

A digital image is the result of a complex *image generation process* that projects a continuous *scene* to a discrete representation. A scene is part of the real world and can refer to any real natural phenomena or describe arbitrary imaginary phenomena that result from human creativity [18, p. 81]. As a result, the image conveys information about the depicted scene, which—by (human) interpretation—translates to a particular *semantic meaning*. The image generation process itself may comprise up to three principal stages, depending on its concrete instantiation, cf. Figure 2.1:

- The image *acquisition* phase is the interface between the real and the digital world, where a scene is projected to a discrete representation. Such projections can take place via an *image acquisition device* that is equipped with a sensor (e. g., a digital camera), or it can be completely software-driven (as in the case of computer-generated images).
- In the *processing* phase, the acquired image is altered (or parts thereof), leading to a processed version. At this stage, we distinguish between image *processing* and image *manipulation*. While processing in general refers to any change of the initial image, a manipulation is

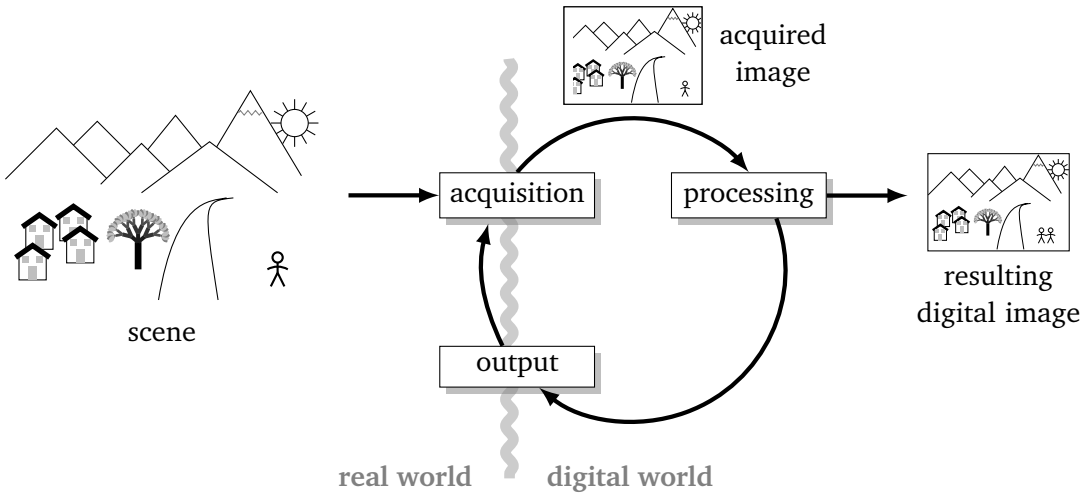


Figure 2.1: A digital image generation process projects parts of the real world (the scene) to a digital representation (the digital image). The acquired image may be post-processed or undergo a re-digitization procedure.

more specific and implies the change of the image's semantic meaning. In other words, a manipulation impairs the *authenticity* of a digital image.

- Finally, the digital image may undergo a re-digitization procedure, where the image is recaptured from the result of an *output* device (for instance photographs of computer displays or scans of printouts). We note that re-digitization may be likewise understood as a special form of post-processing: the resulting digital image is a modified version of the image fed into the output device. We will make us of either perspective, depending on whether the explicit or implicit view is more convenient in a particular situation.

Each of the above steps may be further subdivided into several independent *components*. These components are instantiated specifically to particular (classes of) generation processes, where a *class* subsumes a number of generation processes that share common settings. Different digital cameras models, for instance, may differ in their system of lenses, the type of sensor, or the demosaicing algorithm. Post-processing may comprise resizing or contrast enhancement operations, or combinations thereof.

2.1.2 Trustworthiness of Digital Images

Whenever digital images are understood as a means to convey information, it is important to ensure the *trustworthiness* of this very information. This means in particular that the image has to be *authentic*, i. e., the image has not been manipulated and the depicted scene is a valid representation of the real world. Often however, it is not only the depicted scene that is considered to convey information, but also the image's *origin* and the mere circumstances that led to the respective image.

Example 1 | Trustworthiness. Consider for instance a photograph that is offered to a reputable newspaper, showing a popular political leader handing over a secret document to some dubious lobbyists. The responsible editor may or may not decide to front-page not only depending on whether the image has been tampered with. However, this decision may also depend on

whether the image was taken by a renowned and trusted journalist or was leaked to that journalist by a person from the orbit of an opposing party competing for votes in an upcoming election.

Hence, judging about the trustworthiness of a digital image means to infer the history of that particular image. While the above example illustrated that, in general, the question of trustworthiness is connected to the question *who* operated the devices involved in the image generation process, we will restrict our considerations on the origin of an image to the determination of the actual devices: The image itself cannot reveal the identity of the device operator.[2] Rather, we will understand available information about the involved devices as indications of image origin, which can be further combined with side-information from other sources.

In other words, we are first of all faced with the question what components have made up the image generation process. Based on hypotheses on the image generation process (and additional side-information), we will eventually come to a conclusion to which extent the image shall be considered trustworthy, if at all. In Example 1 above, low image quality due to strong JPEG compression artifacts may hint to a post-processing step and thus diminishes the trustworthiness of the offered image (independent of its actual source). In a similar fashion, the editor may put little trust in the image when she finds out that it was actually captured with a low-cost consumer device, whereas the journalist is known to favor high-end full-frame cameras.

2.1.3 Digital Image Forensics

The primary objective of forensic sciences is the systematic reconstruction of events as well as identification of entities involved. The forensic analysis of digital images (or *digital image forensics*) then refers to the reconstruction of the generation process of a given digital image, where the main focus lies on inference about the image's authenticity and origin. In a strict sense, the term 'forensic' denotes the application of scientific methods to the investigation and prosecution of a crime, i. e., outcomes of a forensic analysis may serve as probative facts in court.[3] Driven by the ubiquity of digital images, the term 'digital image forensics' is used for broader contexts in the literature, with a wide range of applications also beyond the courtroom. Although trustworthiness (and in particular authenticity) is a prerequisite for images being introduced as pieces of evidence to the court [200, i. a.],[4] we expect and demand trustworthiness in *whatever* situation we (have to) rely upon an image. This is to say that applications of digital image forensics have not only a legal but in general also a very strong social dimension.

Throughout this text, we will therefore understand digital image forensics rather general as *any systematic attempt to assess the trustworthiness of a digital image based on its generation process*. The call for systematic approaches emphasizes the link to forensic sciences in general,

2 This is only true unless the image generation process is specifically designed to link biometric identifiers with the resulting image [16], or the image by chance exhibits reflections of the photographer. As to the latter, we refer to Section 5.1 of Ref. [70] for an indicative example.

3 The word has its etymologic roots in the Latin word *forum*, which means 'main square', a place where public court hearings took place in ancient times.

4 Whether digital images are eligible for admission as pieces of evidence in court at all is, at least in Germany, still subject to an ongoing discussion [190, 134].

and consequently, we will refer to a subject that conducts analyses of digital images for the purpose of digital image forensics as a *forensic investigator*.

2.1.4 Image Characteristics and Identifying Traces

The key to digital image forensics is the assumption that components of the image generation process affect characteristics of the resulting image and that these *image characteristics* are specific to the respective (class of) generation process(es). Both because of the very existence of certain components and because of differences in their manufacturing and implementation, different generation processes vary in the way how they impact the characteristics of the resulting image. Eventually, these variations permit inference about particulars of the image generation process and thus allow us to assess the trustworthiness of a given image.

De Rosa et al. [203] reformulate the above assumption as a two-part image decomposition. The first part relates to properties of the scene itself and hence conveys the semantic meaning of the original image. Parameters of the image generation process and their implications on the resulting image determine the second part. Although this decomposition appears rather theoretical (Because a digital image by definition only exists via a particular generation process, it is hard to imagine an "ideal" image that is independent of this very process.), we will adopt the idea behind this construction and henceforth distinguish between *characteristics of the scene* and *characteristics of the generation process*, respectively.[5]

Figure 2.2 extends the previous Figure 2.1 and illustrates the influence of scene and generation process characteristics in the context of digital image forensics. Red circles and black squares indicate, in a stylized manner, that each component of the generation process may leave its very own characteristics (here exemplarily shown for acquisition and processing stage, respectively). Technically, the goal of the forensic investigator is to identify which (class of) generation process(es) led to the image under analysis. It is for this reason that we speak of *identifying traces* when we refer to process-specific characteristics of digital images.

Identifying traces in general possess a high *inter-class similarity* and a low *intra-class similarity*. This means that respective characteristics are highly similar for all images of the same (class of) generation process(es), but differ for images of different (classes of) generation processes [85]. The above identification problem is approached by comparing *reference traces* of known instances of generation processes with traces found in the image under investigation. Digital image forensics is hence a *decision problem*, where the forensic investigator needs to assign a given image of unknown provenance to a known (class of) generation process(es)—or to none of them, when no match is found.

In general, characteristics of different components are not independent of each other. Each individual component of the image generation process can interfere with or even wipe out traces of previous components. This means that characteristics of earlier stages are not necessarily present in the final image anymore. Forensic investigators may assess image authenticity not only based on the *presence* of identifying traces but also explicitly based on the *non-existence* of (expected) traces or, more general, their *inconsistency* by interpreting missing identifying traces

5 However, we do not entirely follow de Rosa et al. [203], who conjecture that the second component is *« a sort of random part of the image »*. While typical generation functions clearly introduce randomness to the image (e. g., temporal sensor noise), it is central to all practical methods in digital image forensics that some deterministic similarities between certain instances of generation processes exist.

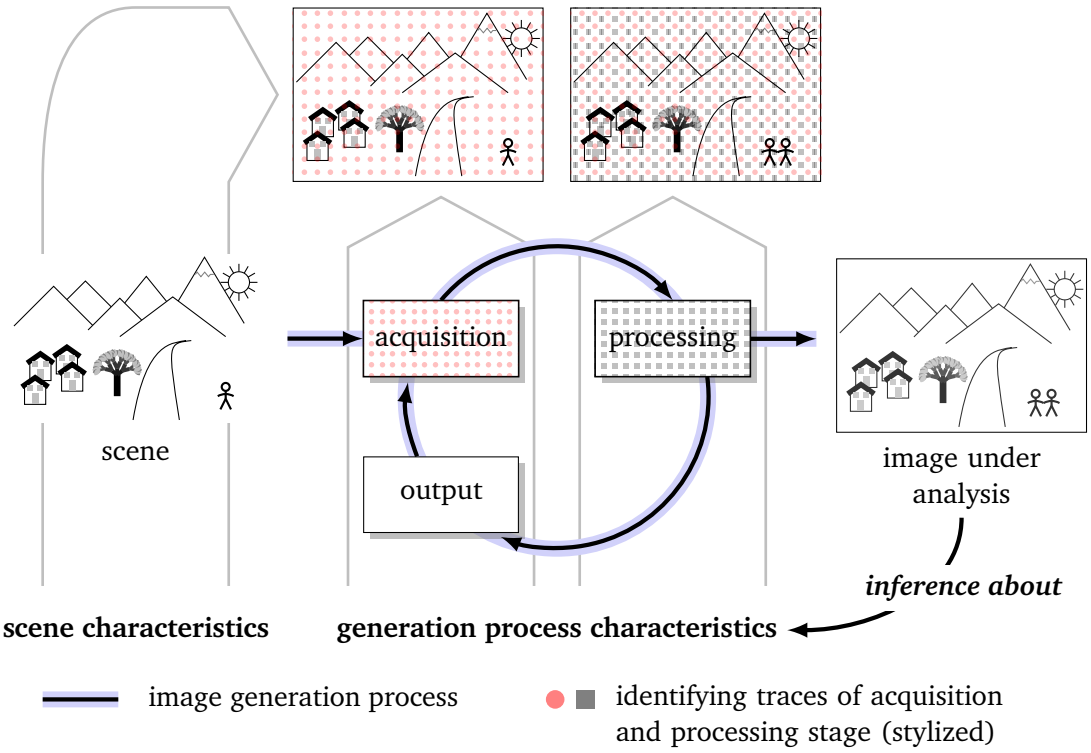
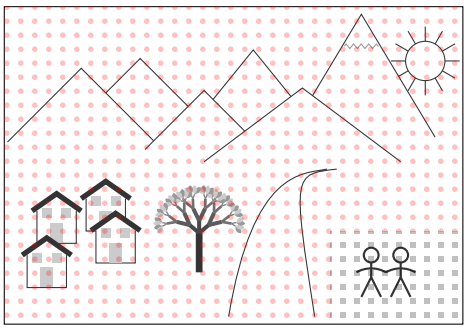


Figure 2.2: Characteristics of digital images divide into scene and generation process characteristics, respectively. Digital image forensics assumes that components of the image generation process leave identifying traces in the resulting digital image. If found in the image under analysis, these traces allow forensic investigators to infer particulars of the generation process.

as strong indication for a post-processed image. Inconsistent here means that traces found in the image under investigation are compatible with reference traces of different, possibly contradictory, generation processes. In Example 1 on page 8 for instance, it may turn out that parts of the image were cropped to hide the presence of other politicians with the goal to change the image's semantic meaning. This leads to an image size incompatible with the sensor size of the camera model as specified in the accompanying Exif data.

More generally, we can distinguish between *inter-characteristic* and *intra-characteristic* inconsistencies. The above mismatch between image resolution and Exif data is a typical representative of the former type, where the analysis of different characteristics leads to contradictory conclusions with respect to the image generation process. Intra-characteristic inconsistencies, on the other hand, refer to spatial variations of one particular characteristic—possibly due to local manipulations of certain regions of the image—that lead to inconsistent conclusions from the analysis of traces in different parts of the image. Figure 2.3 gives an illustrating example, where a local manipulation of the lower-right part of the image led to post-processing traces as well as missing traces of the acquisition phase in the respective image region. This of course implies that the corresponding characteristics can be analyzed locally, i. e., that respective identifying traces are expected to be present (and measurable) throughout different parts of the image. The above example illustrates that this is not a prerequisite for the existence of inter-characteristic inconsistencies: size and metadata are characteristics of the image as a whole.



identifying traces of acquisition and processing stage (stylized)

Figure 2.3: Image characteristics interfere with each other. Local image processing may lead to inconsistent identifying traces (here in the lower right part of the image). See also Figure 2.2.

2.1.5 Counter-Forensics

To justify the additional trust that we place in digital images through forensics, it is important that the limits of image forensics are known and will eventually be overcome. The invocation of an image generation process is typically an intentional act that is carried out by a human being for a specific reason. The resulting image is supposed to convey particular information, either about the depicted scene (and subjects and objects therein) or the image creator itself. As a consequence, there is little doubt that farsighted actors will do their best to influence the result of a potential forensic analysis to a favorable outcome for themselves, i. e., to undergird the information they want the image to convey. In the above Example 1, for instance, the questionable source from the opposite party may want to cover up the fact that the image has actually been manipulated and may also try to conceal the origin of the image.

This makes digital image forensics a *security discipline*, which needs to anticipate *intelligent counterfeiters*, and which has to be measured by its resistance to countermeasures. The research field that challenges digital forensics and systematically explores its limitations against such counterfeiters is called *counter-forensics*, or anti-forensics. Both terms are used synonymously in the literature. We prefer the former because it better reflects the pragmatic reaction to forensics, as opposed to a normative disapproval of forensics.

In a very general setting, counter-forensics has been defined by Harris [95] as any attempt *«to compromise the availability or usefulness of evidence to the forensics process»*. In the context of digital image forensics, counter-forensics translates to the modification or replacement of (parts of) the original image generation process so that the resulting *counterfeit* does not evince sufficient identifying traces of the original generation process. Counterfeiters are however not completely free in their choice of countermeasures, because they need to preserve the intended semantic meaning of the image. Furthermore, it is clear that counterfeiters need to be aware of *new* identifying traces, which may potentially link the counterfeit to the modified image generation process. This sets the stage for an arms race that lets counter-forensics stand in an equally productive relation to image forensics like cryptanalysis to cryptography. Therefore we borrow from the cryptanalysis terminology and call a counter-forensic scheme *attack* (against digital image forensics).

Besides the need to assess and improve the reliability of forensic methods, two more reasons motivate research on counter-forensics. First, generation process characteristics indirectly reveal information about identities of the author of a digital image or depicted subjects therein

(cf. Section 2.1.2). This is not always wanted. Researchers have studied systems providing *unlinkability* and *anonymity* in digital communications for a long time [30, 191]. All these efforts are useless if the connection to the subject can be reestablished by forensic analysis of the message. Hence counter-forensic techniques to suppress traces of origin in digital images are a relevant building block for anonymous image communication, which can be useful in practice to protect the identity of sources, e. g., in the case of legitimate whistle-blowing.

Second, some authors argue that counter-forensics, if implemented in image acquisition devices, can be useful to hide details of the internal imaging pipeline and thus *discourage reverse engineering* [213]. We mention this motivation for completeness, but remain reserved on whether current counter-forensics is ripe enough for this purpose. Our main concern is that counter-forensics often implies a loss of image quality. Camera manufacturers, for instance, compete for quality. It is hard to believe that they would sacrifice a competitive edge for making reverse engineering a little harder. Yet we are curious to see promising applications in the future.

2.2 Variants of Digital Image Forensics

So far, we have considered digital image forensics in a very general setting. In particular, we have not made any assumption with regard to the access the forensic investigator has to components of the image generation process or to their respective inputs and outputs. The literature usually understands digital image forensics in a more narrow sense and refers to *passive–blind image forensics* [181]. Before we follow this established convention in the remainder of this text, we have to delineate the meaning of 'passive' (Section 2.2.1) and 'blind' (Section 2.2.2) in this context, and thereby set up the scope of the following discussions (Section 2.2.3).

2.2.1 Passive vs. Active Image Forensics

Digital image forensics is called *passive* if the forensic investigator cannot interfere with the image generation process and control type and/or appearance of identifying traces. The image generation process is considered as a 'read-only' procedure and the forensic investigator is confined to examine image characteristics that are generated by this process.

Identifying traces in passive image forensics in general divide into *device characteristics* and *processing artifacts*. The former refer to inherent variations between different acquisition (or output) device(s) and thus allow inference about the origin of a given digital image. Such variations may exist, for instance, because manufacturers use different components or adjust parameter settings for different devices. They can also be caused by (unwanted) inherent technological imperfections, such as sensor defects. Processing artifacts, on the other hand, relate to identifying traces that are introduced by post-processing the acquired digital image. Hence, they are a means to assess image authenticity. Similar to device characteristics, different processing procedures may vary in the characteristics of resulting traces and thereby allow inference about the type of processing. Both device characteristics and processing artifacts can be tested for their presence and consistency, whereby inconsistent device characteristics are themselves a processing artifact.

Active approaches differ from passive approaches in that the generation process is purposely modified at an earlier stage to leave behind specific identifying traces. This auxiliary data, would—once being tied to the image—establish a link to the image's origin or ensure the image's authenticity, respectively. Typical instances of active approaches attach metadata to the image (e. g., a cryptographic signature [71] or a robust hash [230]) or embed a digital watermark directly into the image itself [43]. Note that the image generation process in Figure 2.1 and the notion of identifying traces are general enough to cover both, passive *and* active approaches. Consider for instance the embedding of a digital watermark, which is just an additional component to the overall image generation process, specifically conceived to produce identifying traces.[6]

Identifying traces in active image forensics are designed to link the resulting image to its origin, or to be sensitive ('fragile') to (certain types of) image post-processing. By testing for their presence and consistency, these traces allow inference about the responsible component itself as well as subsequent components in the image generation process. In contrast to identifying traces in passive approaches, type and appearance of such traces can be chosen in anticipation of potential subsequent steps of the image generation process. They can also be designed based on cryptographic protocols to guarantee trustworthiness by means of mathematical proofs.[7] Active approaches ideally need to be implemented directly in the acquisition device. It is not possible to infer parts of the generation process prior to the active insertion of respective traces. This reliance on special-purpose hardware is one of the major drawbacks of active image forensics: it does not allow to assess the trustworthiness of arbitrary images of unknown provenance. Moreover, recent examples of hacked active authentication systems of major digital camera vendors suggest that such systems create a deceptive impression of trustworthiness when implementation errors cannot be ruled out.[8]

2.2.2 Blind vs. Non-Blind Image Forensics

Digital image forensics is called *blind* if the forensic investigator is confined to examine the final output of the generation process. In particular, knowledge neither of the original scene nor any intermediate result of the generation process is available at the time of analysis. This includes the forensic investigator's uncertainty about whether the image under analysis has been subject to any kind of post-processing. A blind analysis however not necessarily implies that the forensic investigator does not have any knowledge of or assumptions about potential components of the image generation process. This applies in particular to active image forensics, because the image generation process has been purposely modified by the investigator at an earlier stage. But also passive forensic investigators may rely on such information. For instance,

6 The distinction between passive and active techniques not completely coincides with the notion of *intrinsic* and *extrinsic* identifying traces, which have been defined by Swaminathan [217, pp. 11 & 13] in a very similar framework as *«traces that are left behind in a digital image when it goes though various processing blocks in the information processing chain»* and *«external signals added to the image [...] after capture»*, respectively. Adding noise to a manipulated image with the goal to let it appear more natural would be extrinsic, just like embedding a watermark.

7 Cryptographic techniques not only allow to authenticate the image and link it to an acquisition device, but can also ensure that no images have been deleted after acquisition [119].

8 Canon's "Original Data Security Kit", for instance, uses the very same cryptographic key for signing images taken with arbitrary devices of the same camera model. Once this key is known (it is stored on the device), it is possible to authenticate arbitrary manipulated images to let them appear trustworthy [209].

it is often known (with reasonably high certainty), which device captured the image when it is analyzed for inconsistent device characteristics and potential manipulations [155, i. a.].

Additional information about intermediate results helps *non–blind* forensic investigators to disentangle scene and generation process characteristics, respectively, and hence to make more informed decisions. Such data may be available from alternative sources (for instance, earlier versions of a processed image that have been published elsewhere), or could have been stored purposely in advance (most likely the acquired image). Side information about the original scene may also be retrieved from other (trustworthy) images that exist of the same scene. Non–blind approaches in general have the advantage to mitigate some of the forensic investigator's uncertainty. At the same time, they are often unviable in practical settings. In particular, non–blind forensics precludes the examination of arbitrary images of unknown provenance.

Swaminathan et al. [220] further divide non–blind forensics into semi–intrusive and intrusive approaches.[9] A *semi–intrusive* analysis considers the image generation process as a black box, which is fed known input signals and allows inference about the overall input-output relation. In an *intrusive* analysis, the forensic investigator also has knowledge of the inputs and outputs of individual components of the image generation process and aims to determine their specific instantiation. This distinction becomes particularly relevant when the forensic investigator is less interested in assessing the trustworthiness of an image, but really in "reverse-engineering" parameters of the image generation process (to which she is granted access). It is then possible to design special inputs, either to the complete generation process or to individual components, that are particularly suitable to analyze and tell apart identifying traces of different instantiations [221].[10]

2.2.3 Passive–Blind Image Forensics

In the remainder of this text, we focus on *passive–blind image forensics* [181] and use this term synonymously with digital image forensics. It follows from the above discussions that passive–blind forensic investigators neither can access or control components of the image generation process (passive) nor have knowledge of its inputs or intermediate results (blind).[11] Their analysis is solely based on inherent device characteristics and processing artifacts, which they (aim to) extract from the image under investigation to infer particulars of the image generation process.

Because the forensic investigator's view is restricted to the image under analysis, passive–blind image forensics will always remain an inexact science. With digital images being projections of the infinite set of all conceivable scenes, the forensic investigator can never fully know whether a depicted scene is indeed a valid representation of the real world, or whether the image has been manipulated in some way. In general, there will always exist a (possibly infinitesimal) residual probability that a particular scene imposes certain image characteristics that appear to the forensic investigator as device characteristic or processing artifact and thus lead to false decisions. Moreover, any image generation process inevitably involves quantization. By definition, quantization causes information loss and hence leaves the forensic investigator

9 In this terminology, blind image forensics is referred to as *non–intrusive* forensics.

10 In an earlier work, Khanna et al. [121] called these specifically designed inputs *probe signals*.

11 Some authors [161, i. a.] only use the term 'blind forensics' instead and do not make the passive character explicit.

with an additional source of uncertainty. As a consequence, any result of passive–blind image forensics has to be understood as indication, which *per se* excludes absolute conclusions.

However, the advantage of passive–blind image forensics is its universal applicability. In particular, it is applicable to the analysis of arbitrary images of unknown provenance. Passive–blind image forensics does not rely on a closed infrastructure of trustworthy special-purpose acquisition devices that actively introduce identifying traces to the resulting images. While this might be a viable option for relatively small-scale applications, it is certainly too costly and most likely politically unviable to be implemented in typical consumer devices.

2.3 Abstraction Layers in Passive–Blind Image Forensics

Passive–blind forensic investigators may examine identifying traces at different 'levels' of abstraction [147, 231]. Such *abstraction layers*—a general concept to the forensic analysis of any type of digital evidence with a semantic meaning [27]—help to encapsulate and study basic properties of scene and generation process characteristics, respectively. Figure 2.4 illustrates relevant abstraction layers in the context of passive–blind image forensics and serves as a blueprint for the following discussions. Levels of analysis and respective image characteristics of interest are arranged row-wise in the upper part of the figure. Dashed arrows indicate how relevant characteristics 'propagate', i. e., which characteristics may interfere with each other. At the lowest level, forensic investigators analyze generation process characteristics by interpreting digital image data as a sequence of discrete symbols without taking any semantic information into account (Section 2.3.1). Higher levels abstract from the plain signal and examine how scene characteristics in general and the image's semantic meaning in particular have been affected by the image generation process (Section 2.3.2). Apart from the examination of image characteristics and scene properties, forensic investigators can further exploit auxiliary metadata that is stored with the image (Section 2.3.3).

2.3.1 Signal-Based Analysis

Signal-based forensic analyses understand the image under investigation as a plain sequence of discrete symbols and ignore its semantic meaning. Both device characteristics and processing artifacts can affect the appearance of this very signal, whereas post-processing may also result in missing or inconsistent device characteristics and *vice versa*[12] (cf. Section 2.1.4). Because the image is analyzed independent of its semantic meaning, ideal identifying traces at the signal-level can only exist if they are independent of the image content [7, 125]. This is also reflected in Figure 2.4, which illustrates that low-level forensic analyses do not take higher-level scene properties into account.

In practice, we observe a continuum of possible analyses from the signal-level to higher levels. First, certain procedures in the image generation process are inherently content-adaptive (for instance edge-directed demosaicing [29] or changes of image resolution based on seam-carving [8]) and forensic investigators may incorporate this knowledge [220, i. a.]. Second, certain parts of an image need to be excluded from the forensic analysis or require

12 Re-digitization may affect identifying traces of previous processing.

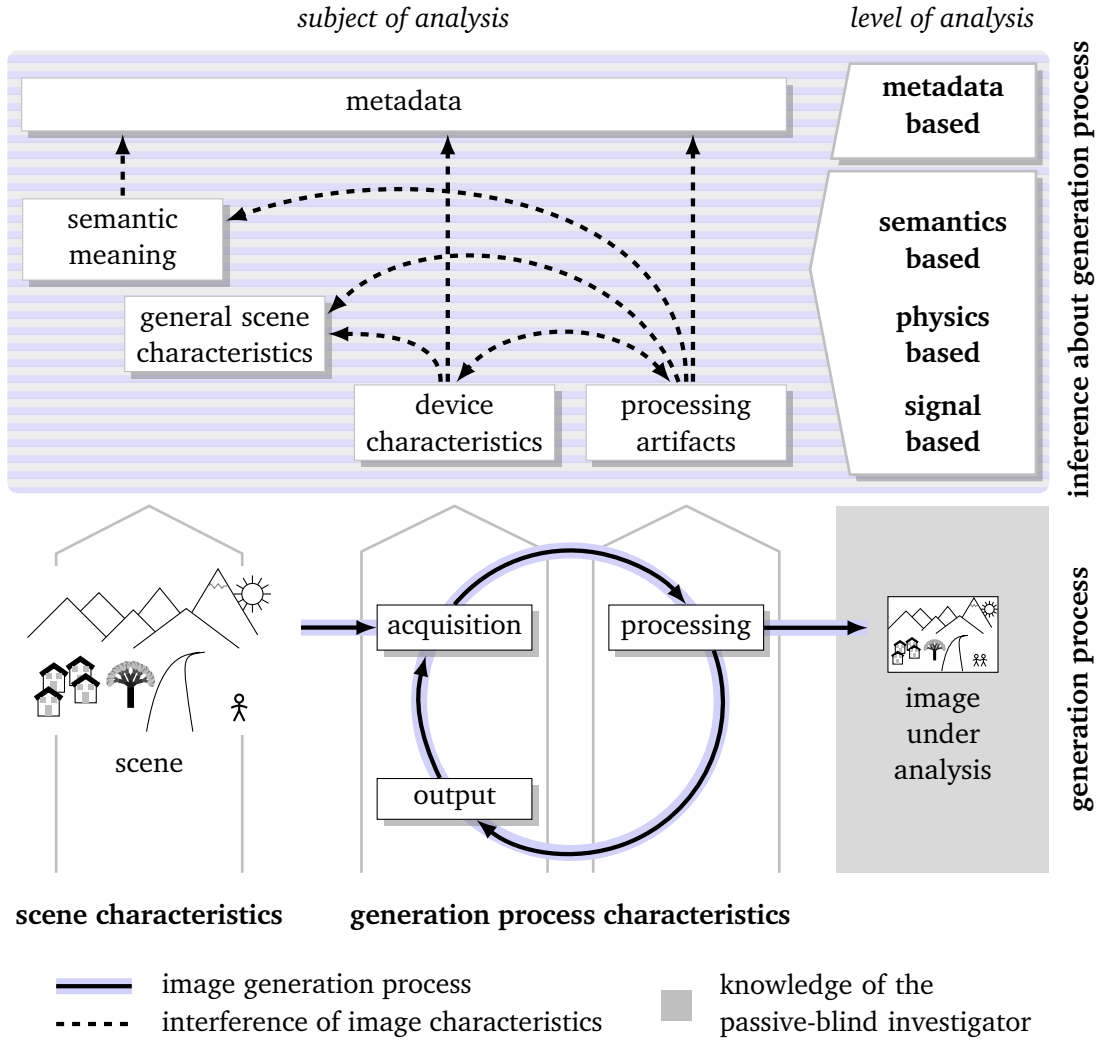


Figure 2.4: Abstraction layers in passive–blind image forensics. For a given image under analysis, passive–blind image forensics infers particulars of the unknown image generation process based on device characteristics and processing artifacts. These generation process characteristics not only interfere with each other but also with characteristics of the scene and auxiliary metadata stored with the image, respectively. Forensic investigators can therefore conduct their analysis at different levels of abstraction.

a special treatment (e. g., saturated or largely homogenous regions in the detection of copy-move forgeries [69, 195]). Finally, signal-level indications may be combined with semantic information because low-level analyses are often not sufficient to answer the general question whether the image under investigation is authentic. More specifically, the sole detection of post-processing in general does not reveal whether the semantic meaning of the image has changed. Therefore, most practical forensic algorithms—although fully automatable in theory—today require a human interpretation of their results.

2.3.2 Scene-Based Analysis

By explicitly taking scene properties into account, forensic investigators abstract from the discrete image data and examine identifying traces at a higher level, cf. Figure 2.4. Here, we can further distinguish between *physics-based* and *semantics-based* analysis. The former relates to general scene characteristics that are inherent to (projections of) the real world, whereas the latter refers to the semantic meaning of the particular image under investigation.

2.3.2.1 Physics-Based Analysis

The premise to any physics-based [180] analysis is that (projections of) real-world phenomena obey certain general physical rules, which are induced by laws of nature. Any violation of these rules is understood as a processing artifact and can serve as identifying trace in a forensic investigation. Ng and Chang [177] called this the 'natural scene quality' of a digital image and mentioned relevant characteristics like illumination, shadows and reflections of objects in the image. Other traces result from geometric constraints [57], (e. g., the relative position of objects in the real world), or texture and surface characteristics.[13] As indicated by Figure 2.4, also device characteristics influence certain general scene characteristics. A typical example are illumination and shadows in an image. They may vary depending on whether a flash was fired or not. Lens radial distortion is another example [37]. Here, genuinely straight lines in the real world appear curved in the depicted scene.

All the above characteristics have in common that they can be studied without explicitly knowing the scene's semantic meaning. It is rather the 'syntax' of the scene that is of interest here. Yet because these characteristics are examined based on the digital image data, their analysis can be conducted by computational methods. Many forensic algorithms of this kind are nevertheless semi-automatic in the sense that they require a human pre-selection of salient objects to analyze [112, 149] and/or human interpretation of the results [201].

2.3.2.2 Semantics-Based Analysis

At the highest level, a semantics-based analysis directly relates to the semantic meaning of the image under investigation (cf. Figure 2.4). The key question to answer is whether the information conveyed in the image is plausible, i. e., whether the image is a valid representation of the real world. This entails the interpretation of the depicted scene and thus requires contextual knowledge, which relates the digital image to the presumed original scene in the real word (for instance, what is depicted in the image, where and when was the image captured, ...). Inconsistencies between context information and semantic meaning of the image are then understood as processing artifact. Because computer algorithms for fully automated and context-dependent image understanding do not exist, a truly semantic image analysis remains a task that involves substantial human interaction. We will not discuss analyses at this level in further detail in the scope of this text.

13 Such tests are not limited to the examination of objects (which have a certain semantic meaning) but may also include surface texture or halftoning pattern of scanned documents [206].

2.3.3 Metadata-Based Analysis

Apart from the image data itself, forensic investigators may exploit the rich source of auxiliary digital data, which typically accompanies the image under investigation. Today, the preferred method to organize and store such *metadata* is specified in the Exif standard [110]. In a forensic analysis, metadata may provide information about the acquisition device and its settings, the time and place of acquisition, the software that was used for post-processing, and much more [4]. (This is also reflected in Figure 2.4, where the metadata field covers the whole range from scene characteristics to generation process characteristics.) Moreover, many digital images are accompanied by low-resolution *thumbnail images* to speed up the preview in file browser applications.

Format, content, and structure of metadata and thumbnail images can serve as a source of forensic evidence on their own. These characteristics allow to assess image trustworthiness independent of the actual image data [4, 207, 118]. It is also possible to link this information to the image and to interpret scene-level and signal-level inconsistencies as processing artifact. In a signal-level analysis, Kee and Farid [117] for instance estimated parameters of the thumbnail generation process from the image data. Semantic inconsistencies can be uncovered by a high-level comparison of thumbnail and full-resolution image [168] or by tests for the plausibility of certain camera settings (e. g., focal length, shutter speed, date and time of acquisition, . . .).

However, a downside of metadata is that it is relatively easy to modify. Because the metadata is stored detached from the actual image, it is always possible to alter parts of or completely replace metadata and thumbnail images in order to conceal particulars of the image generation process. As a result, the forensic analysis of auxiliary digital data in most cases can only be a first step to a detailed image-data based analysis.

2.4 Specific Goals of Passive–Blind Image Forensics

Passive–blind assessment of image trustworthiness subsumes a variety of more specific goals. Figure 2.5 illustrates that we can in general distinguish three categories. Tests for the very *presence* (or absence) of components of the image generation process at the acquisition, post-processing or output phase allow to detect computer-generated, processed, or recaptured images, respectively (Section 2.4.1). By examining the concrete *instantiation* of certain components, it is further possible to identify the source of an image, to infer its processing history, or to determine the output device in a re-digitization procedure (Section 2.4.2). Finally, the *linkage* of a number of image generation processes can be of interest, for instance with respect to the time of their invocation or their equivalence (Section 2.4.3).

2.4.1 Tests for the Presence of Components

2.4.1.1 Detection of Computer-Generated Images

The detection of computer-generated images belongs to the first problems studied in the area of digital image forensics [60, 177].[14] The main difference between computer-generated

14 It can be even traced back to much earlier applications in content retrieval [6].

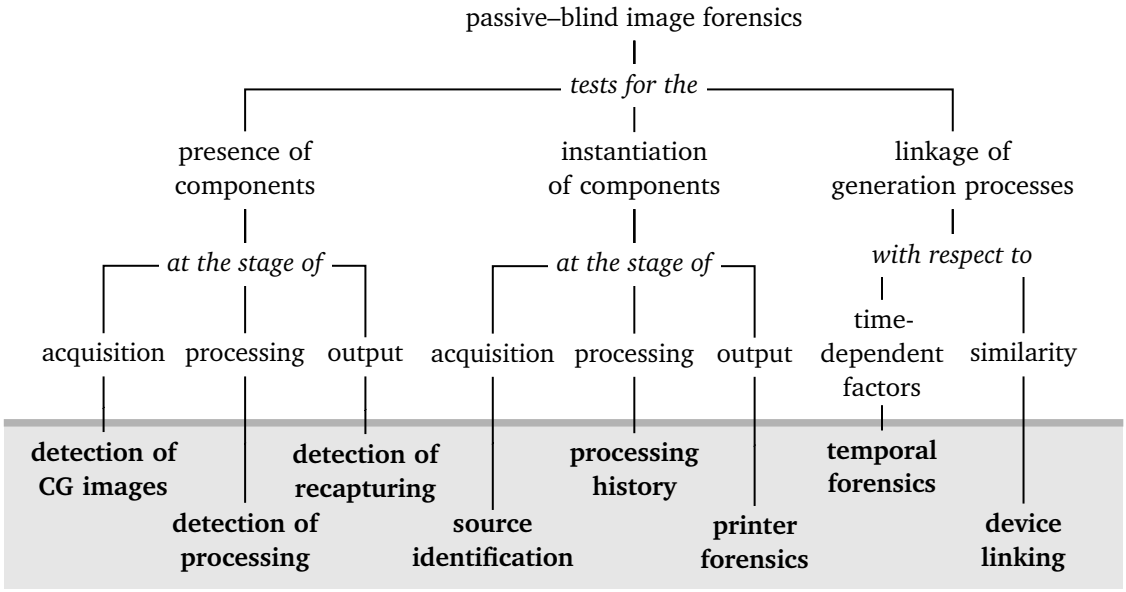


Figure 2.5: Specific goals of passive–blind image forensics.

images and other types of digital images is the absence of an image acquisition device in the image generation process. Rather, computer-generated images are digital representations of imaginary scenes that result from human creativity. It is this lack of a real natural phenomenon in combination with the absence of an acquisition device that has motivated virtually all existing detectors of computer-generated imagery.

More specifically, computer-generated images can be detected by the physics-based analysis of general scene characteristics, where the working assumption is that projections of complex real natural phenomena are hard to synthesize with a computer. Image rendering software is confined to simplified models of reality, which need to be kept computationally tractable [180].[15] More general, it is believed that digital images of natural phenomena exhibit common statistical properties independent of the actual acquisition device and the depicted scene. These *natural image statistics* include for instance empirical measures such as power law and scale invariance [62], or spatial and inter-scale correlation [21]. Computer-generated images do not fully conform to these characteristics and thus become detectable [177, 159]. At the signal-level, image rendering software typically does not mimic the exact processing chain of imaging devices, making computer-generated images detectable because of missing device characteristics [47, 75].

2.4.1.2 Detection of Processed Images

One of the most fundamental problems in digital image forensics is to infer whether the image under analysis has been modified after initial acquisition. At the signal-level, processing artifacts can be detected by testing either for the existence of processing-specific characteristics [69, 196, 72, 208, 131, 211, i. a.] or for missing or inconsistent device characteristics. As to

15 Ng et al. [180] point out that the most severe simplifications occur at the level of object geometry, surface, and illumination of the depicted scene. Other authors further note differences in the color representation [6, 237].

the latter, the literature has so far been mostly concerned about the examination of traces generated in the acquisition phase [147, 198, 154, 111, i. a.] and only few works have focused on characteristics of the output device [116]. Inconsistencies can occur both inter- and intra-characteristic and may become particularly indicative when parts of different images are combined ('spliced' [176]) into one image. The advantage of device characteristics over processing artifacts is that investigations are in general not restricted to identifying traces of specific post-processing procedures. At the same time, however, device characteristics are only of limited use for inference about the complete processing history, cf. Section 2.4.2.2.

Processing artifacts can also be detected at the scene level by checking for consistent general scene characteristics. However, scene-level analyses today typically require a considerably higher degree of human interaction (and interpretation) than signal-level analyses (cf. Section 2.3.2.1). The latter, because of their high practicability and relative ease of automation, have traditionally received more interest. Yet scene-level analyses offer some additional advantages to the forensic investigator. First, it is in general a challenging task to create image manipulations that take into account and correct for the plethora of complex scene characteristics. Second, scene-based analyses work at a higher abstraction layer and are often closely tied to the examination of (projections of) individual objects. This results in a certain degree of invariance to post-manipulation reduction of image resolution and/or quantization—or, more general, any information-reducing operation, which will inherently affect subtle identifying traces at the signal level.

2.4.1.3 Detection of Recaptured Images

Images that are re-digitized from two-dimensional (planar) products of an output device (e. g., printouts) substantially differ from direct projections of real-world, natural phenomena. As a consequence, *recaptured images* are not only detectable at the signal-level by testing for characteristics of the output device [141], but also by the analysis of general scene characteristcs. The latter relate to general properties of printed [60, 242, 76] or displayed [26] images (e. g., surface texture pattern) and make the detection of recaptured images an exception in the sense that already the presence of certain scene properties is considered as identifying trace of the respective image generation process. This inconsistency in terminology can be avoided by interpreting re-digitization as a form of post-processing (cf. Section 2.1.1) and by understanding these traces as a special form of processing artifact.

2.4.2 Tests for the Instantiation of Components

2.4.2.1 Device Identification

The goal of *device identification* is to link the image under investigation to a particular (class of) acquisition or output device(s). The former problem is typically referred to as *source identification* [125], whereas it is mostly the problem of *printer forensics* that has received interest on the output side. By its very definition, device identification is in general approached by signal-based or physics-based analyses of appropriate device characteristics. We note however that printer forensics, in the way it is discussed in the literature [1, 2], is not always a truly blind problem. The forensic investigator is usually assumed to be in possession of the analog output and has full control of the subsequent re-digitalization step (typically high-resolution scanning). This is more a problem of *computational* [64] forensics, where the

re-projection to the digital domain solely serves as a means to support forensic analyses.[16] Therefore, we refer to Chiang et al. [35] for an extensive review of relevant approaches to printer identification and focus on source identification in the following.

Although of main interest in theory, the identification of the very device [79, 155, 48] is not always possible for practical reasons. Apart from situations where no suitable identifying traces are known, all existing methods require access to the specific device that is to be identified, or at least to a number of reference images acquired with that device. If these prerequisites are not met (i. e., no reference traces of the device are available), the forensic investigator may resort to lower levels of *identification granularity*, such as device model [125, 11, 220, 28, 85], brand [240] or type (e. g., consumer-camera vs. signal-reflex camera, or digital camera vs. flatbed scanner) [123, 164, 53], cf. Figure 2.6. In contrast to device identification, low-granularity approaches only require reference traces of devices that are representative of a particular (class of) image generation process(es) at the given level of granularity. This way, the forensic investigator can infer at least some information about the image origin. Empirical results from the literature however indicate that no single device characteristic is known to distinguish between arbitrary low-granularity classes of acquisition devices. Practical settings may hence require a combination of several device characteristics to narrow down the specific device as best as possible.

2.4.2.2 Processing History

Apart from the question *if* the image under analysis has been modified (cf. Section 2.4.1.2), it is also of interest to know *how* it has been processed. Such information is particularly useful in the assessment of image authenticity. Inference on the *processing history* of digital images involves to determine the sequence of post-processing steps that led to the final image, as well as the parametrization of the respective components [223]. This is mostly a problem at the signal-level, because scene-based analyses can only provide high-level indications (for instance, which objects/parts of the image have been modified).

If forensic algorithms for the detection of processed images exploit identifying traces of a specific type of processing, they inherently provide some information about the image's processing history. Nevertheless, a full recovery of the processing history is typically much more involved, because identifying traces of earlier processing steps may be altered or attenuated by subsequent processing steps. A further, more specific issue arises from ambiguous processing artifacts that allow to identify a particular processing component, but not its parametrization. For instance, several resizing parameters are known to introduce equivalent artifacts [197, 129], or may be hardly distinguishable from JPEG post-compression [74].

We stress once more that the detection of arbitrary image processing and intentional image manipulations not necessarily coincide. Ultimately, forensic investigators need to distinguish between *legitimate* and *illegitimate* post-processing in order to assess image authenticity. While any type of processing may call into question the image's authenticity, it requires interpretation and understanding if and how the detected processing has affected the image's semantic meaning. Although of high practical relevance, the literature has been very indefinite in this regard. Terms like 'forgery', 'tampering' and 'manipulation' have been used more or less

16 Other non–blind examples are given by Gaubatz and Simske [77, 78], who discuss combined printer-scanner identification against the backdrop of highly structured and computer-generated images at the input side.

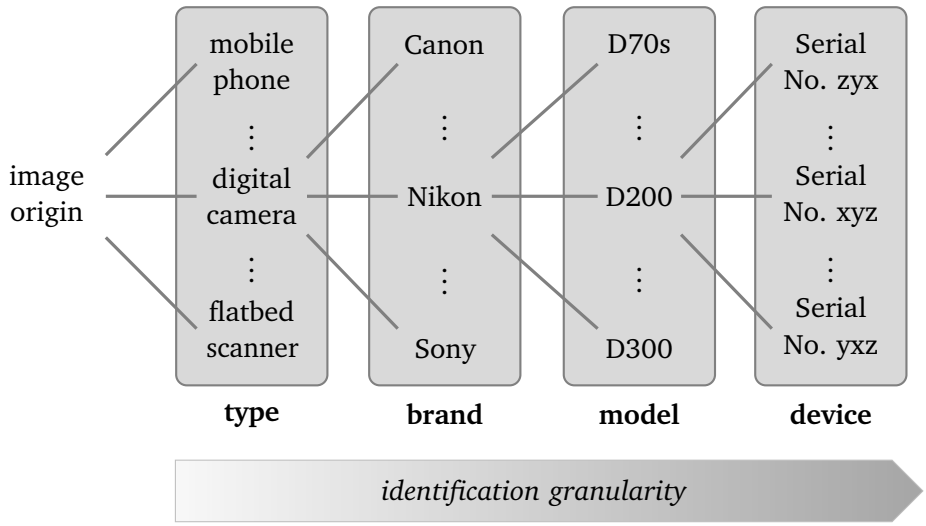


Figure 2.6: Levels of identification granularity in source identification.

synonymously, sometimes amplified by the label 'malicious' [69, 196] to better distinguish between image processing and image manipulation. We refer to Chapter 3 for an attempt to incorporate the notion of authenticity into a general passive–blind forensics framework.

2.4.3 Tests for the Linkage of Generation Processes

Apart from the analysis of single images, the examination of relations between a number of digital images with respect to their generation processes can be of interest. The most relevant questions are whether a group of images have been generated using the same input or output device, or in which temporal order a number of given images have been generated.

2.4.3.1 Device Linking

Tests for a common input or output device are known as *device linking* [87]. Similar to device identification, it is studied under the premise that images of the same provenance will exhibit common device characteristics at the signal-level. The scenario differs from plain device identification in that reference traces of the set of potential devices need not to be known. The problem is approached by extracting suitable device characteristics from one (or several) images, which are then tested against all remaining images. Device linking is in general the harder task: Reference data in source identification is usually more reliable (of higher quality) because it is obtained under controllable conditions.[17]

2.4.3.2 Temporal Forensics

The objective of *temporal forensics* [162] is to establish an order of a number of images—typically acquired with the same device—with respect to the time of their generation. More specifically, we can distinguish two scenarios depending on whether the images have been

17 If reference traces are aggregated from sample images, source identification and device linking converge, when the number of images of the same (unknown) device is large enough [15].

Table 2.1: Selected goals of passive–blind image forensics and relevant abstraction layers.

level of analysis	CG detection	detection of processing	detection of recapturing	source identification	processing history	printer forensics	temporal forensics	device linking
semantics	(×)	×			(×)		(×)	
physics	×	×	×	×	×	×	×	×
signal	×	×	×	×	×	×	×	×
metadata	×	×	×	×	×		×	×

generated independent of each other or whether all images are assumed processed versions of each other. A necessary condition for the first scenario is that certain device characteristics evolve over time, and that corresponding reference traces of at least one point in time are available. As pointed out by Fridrich and Goljan [67], suitable identifying traces are most likely due to undesired device imperfections and wear effects that are not part of the generation process by design. The second scenario is closely related to the determination of processing history [203, 45]. To a certain degree this problem is non–blind because the forensic investigator observes intermediate results of the image generation process, however without knowing how they relate to each other.

Mind that known signal-based or physics-based analyses cannot determine the *exact* time of image generation, but rather hint to a possible time interval. Exact determination is only possible with additional side-information about the real world (either because of semantic information or subtle traces of natural phenomena similar to the electric network frequency [94] in digital video and audio analysis), or metadata support.

2.4.4 Summary

Table 2.1 summarizes the two preceding sections and lists the various specific goals of passive–blind image forensics together with respective relevant levels of analysis. For completeness, we note that metadata may accompany signal- and physics-based analyses in virtually all of the above discussed problems with the exception of printer forensics.[18] A pure semantic analysis, on the other hand, is only viable with regard to the detection of processing artifacts. In special cases, it may also expose computer-generated images because of unrealistic content. Or it may reveal information about the time of acquisition (e.g., when a calendar or watch is depicted).

2.5 Specific Goals of Counter-Forensics

Given Harris' [95] broad definition of counter-forensics (cf. Section 2.1.5), it is clear that counter-forensic attacks are in general relevant to each of the above-discussed specific problems in digital image forensics. However, it is not our intention to reiterate every detail from the counterfeiter's perspective. More general, the counterfeiter ultimately strives for counterfeits

18 It is reasonable to assume that metadata of the output device does not exist.

Table 2.2: Specific goals of counter-forensics.

	suppression	**synthesis**
device characteristics	hide image origin	forge image origin; remove processing-related inconsistencies
processing artifacts	hide traces of processing	remove processing-related inconsistencies

that appear trustworthy, i. e., counterfeits, which do not raise the forensic investigator's suspicion. This boils down to two principal goals a counterfeiter may pursue, namely the

- *suppression of identifying traces*, or the
- *synthesis of artificial identifying traces*, respectively.

Table 2.2 illustrates that both, suppression and synthesis can relate to either device characteristics (Section 2.5.1) or processing artifacts (Section 2.5.2).

2.5.1 Suppression and Synthesis of Device Characteristics

Device characteristics allow forensic investigators to link the image under investigation to a specific device or a class of devices. Counterfeiters, who want to *hide the origin of an image*, will hence attempt to suppress relevant device characteristics so that a forensic analysis of the image does not reveal which device(s) were part of the image generation process [83, 204, 142, 167]. Attacks against device identification schemes may further aim to *forge the origin of an image*, i. e., to let the image appear as if it was generated using a device different from the original one. Here, the counterfeiter not only needs to suppress traces of the true device, but also needs to synthesize traces of the *target device* [83, 130, 214, 142, 215]. This type of attack is typically considered more severe, because it can lead to a false accusation with respect to the owner of the pretended device.

Synthesis of device characteristics is also a building block to the creation of plausible image manipulations. In anticipation of forensic investigators that test for the consistent presence of certain device characteristics, counterfeiters have to *remove processing-related inconsistencies* therein. They do so by synthesizing relevant device characteristics, which either relate to the original or to completely new device(s).

2.5.2 Suppression and Synthesis of Processing Artifacts

Counter-forensic attacks with the goal to *hide traces of (post-)processing* not only need to make sure that device characteristics appear consistent but will have to suppress processing artifacts in general [128, 129, 213, 24]. Often, this rather strict requirement is relaxed by "replacing" conspicuous with plausible processing artifacts, for instance by exploiting lossy JPEG compression. This way, subtle yet telling traces of previous manipulations are wiped out, whereas established habits suggest that rather obvious JPEG artifacts are not *per se* considered critical in many situations.

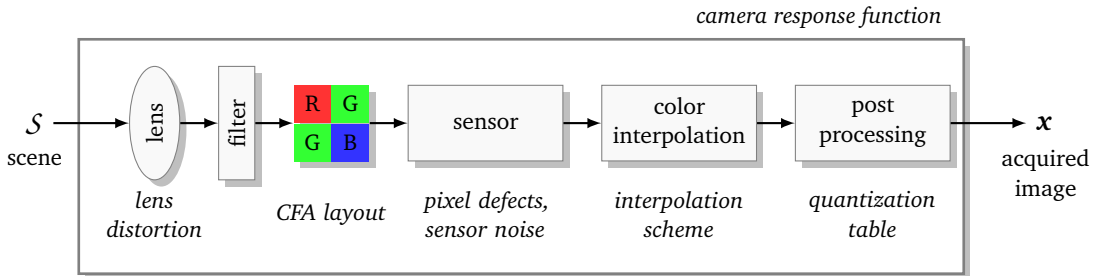


Figure 2.7: Digital camera processing pipeline. Relevant identifying traces are printed in italics.

We note that there also exist situations where the synthesis of processing artifacts is viable. Consider, for example, a manipulation where only parts of the image were resized to change the image's semantic meaning. Because resizing itself—if applied to the whole image—may not impair authenticity, a counter-forensic attack could attempt to synthesize traces of resizing in the remaining parts of the image (without actually resizing these parts). By *removing* such *processing-related inconsistencies*, the whole image consistently appears resized and thus hides the manipulation.

2.6 Selected Identifying Traces

We close this chapter with a brief informal overview of some of the most relevant device characteristics (Section 2.6.1) and processing artifacts (Section 2.6.2) that have found application in the literature. The intention of this section is not to give an in-depth exposition of concrete forensic algorithms, but rather to discuss general properties of identifying traces these algorithms may exploit. This way, this section serves as a connecter to the following Chapter 3, where we turn to a formal perspective on identifying traces and digital image forensics in general.

2.6.1 Device Characteristics

Components of acquisition and output devices, respectively, leave behind characteristic traces in the resulting image. By testing for presence and consistency of these device characteristics in a given image under analysis, forensic investigators infer details about the image origin and its authenticity. With reference to Section 2.4.2.1, we focus on characteristics of the acquisition device, and in particular on digital cameras. Given their widespread use in our everyday life, this type of acquisition device has by far received the most interest in the literature. Nevertheless, many of the device characteristics we will discuss in the following also have their place in the forensic examination of other types of acquisition devices, most prominently flatbed scanners. We will mention relevant differences whenever appropriate. We refer interested readers to Chiang et al. [35] for a comprehensive review of scanner characteristics.

Figure 2.7 depicts a stylized processing pipeline of a typical digital camera with its most relevant components [199, 100]. The incoming light of the scene is focused on the *sensor* by a *system of lenses*. An *optical filter* is interposed between these components to reduce undesired

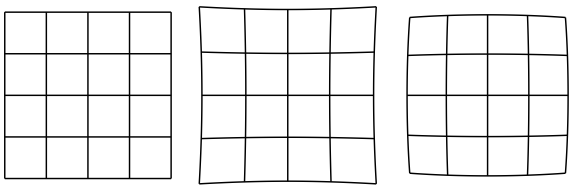


Figure 2.8: Lens radial distortion lets initially straight lines appear curved in the image. From left to right: an ideal rectangular grid, and the same grid subject to barrel and pincushion distortion, respectively.

light components (e. g., infrared light). Typical digital imaging sensors capture image pixels by distinct CCD or CMOS sensor elements that capture the incoming light and output an electric charge proportional to the light at that location. Although digital camera images usually consist of the three color channels red, green and blue (RGB), the sensor itself is color-blind as it can only measure the light intensity. To obtain a color image, the vast majority of digital camera designs employ a *color filter array* (CFA), such that each sensor element only records light of a certain range of wavelengths. The remaining color information has then to be estimated from surrounding pixels of the raw image (after possible pre-processing and white-balancing). This process is usually referred to as *color filter array interpolation* or *demosaicing*. After CFA interpolation, the image is subject to a number of camera-internal post-processing steps, including for instance color correction, edge enhancement, and finally compression.

2.6.1.1 Lens Distortion

Each digital camera is equipped with a complex optical system that projects a scene to a sensor of much smaller dimension. This projection is in general not perfect and a plethora of *lens distortions* (or *aberrations*) can be found in digital images despite extensive compensation efforts of lens and camera manufactures.

Because different camera models employ different optical systems, which all have their own individual distortion profile, these aberrations may serve as identifying traces of the specific camera model.[19] Choi et al. [37] first took this path and proposed to exploit *lens radial distortion* in a forensic context. Lens radial distortion is a non-linear geometric distortion that lets initially straight lines appear curved. In general, we can distinguish between barrel and pincushion distortion, respectively (cf. Figure 2.8), whereas shape and strength depend on the concrete lens(es) in use.

Van et al. [229] made similar endeavors with respect to the analysis of *chromatic aberrations*, which occur because of lens-specific variations in the dispersion index for light components of different wavelengths. This, by Snell's law, causes a polychromatic ray of light to be spread over different positions of the sensor plane, cf. Figure 2.9. Here, axial chromatic aberration refers to wavelength-dependent longitudinal variations of the focal point along the optical axis, whereas lateral chromatic aberration explains off-axis displacements of different light components relative to each other. Although often ignored by the viewer, in particular lateral chromatic aberrations are a visible part of almost any digital camera image, where they manifest in color fringes along edges.

Vignetting is another type of visible lens distortion that describes the radial decrease of light intensity towards the corners of an image. As with chromatic aberrations, its appearance is

19 Yu et al. [243], against the backdrop of a small number of shots of one very specific probe image, report that chromatic aberrations might even be specific to *individual* lenses. Yet it seems too early to draw general conclusions and further experiments are necessary.

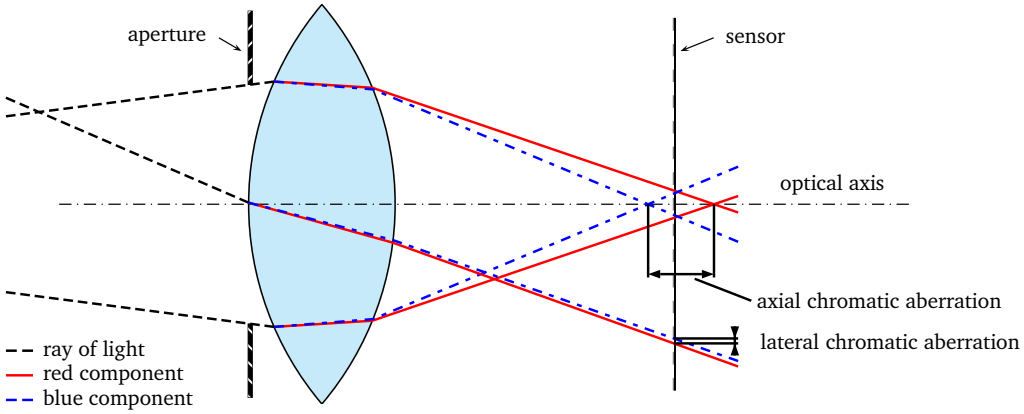


Figure 2.9: Formation of axial and lateral chromatic aberrations due to variations in the lens' dispersion index for light components of different wavelengths. The figure is taken from [86] and represents the complex optical system of real digital cameras in a stylized manner as a single lens.

believed to be characteristic to the optical system employed by the camera model and thus gives rise to applications in source identification [158].

A general influencing factor to the appearance and strength of most lens distortions throughout different regions of an image is the (radial) distance to the *optical center* of the image.[20] Hence, as first detailed by Johnson and Farid [111] for the specific example of chromatic aberration, image manipulations may introduce intra-characteristic inconsistencies when portions of an image are replaced by image parts with a mismatching local distortion profile.[21] Figure 2.10 gives a concrete example and illustrates how lateral chromatic aberrations vary according to the position in the image. Very recently, Yerushalmy and Hel-Or [241] extended this observation to so-called *purple fringing aberrations* (FPA), which are attributed (amongst others) to micro-lenses in front of individual sensor elements.

2.6.1.2 Sensor Imperfections

Not only the optical system bears imperfections, but so does the sensor in the process of converting incoming light to a digital image. *Sensor imperfections* caused by variations in the manufacturing process and sensor wear-out have been on the agenda of forensic investigators from the very beginning.

More specifically, any sensor introduces a certain amount of *noise* to the resulting image—slight fluctuations in the intensity of individual pixels even when the sensor plane was lit absolutely homogeneously. Sensor noise broadly classifies into temporal noise and spatial noise. Noise components that are stochastically independent over different invocations of the image generation process belong to the former type, with shot noise and read-out noise being typical representatives. On the contrary, spatial noise is relatively stable over time and

20 The optical center is the point where the optical axis strikes the sensor plane and does not necessarily coincide with the geometrical center of the digital image. As such, it can be considered a device characteristic on its own.

21 Empirical results by Gloe et al. [86], however, suggest that complex lens systems often lead to unpredictable (or perhaps: not yet fully understood) device characteristics that may be easily mistaken for processing artifacts.

Figure 2.10: Lateral chromatic aberration in a digital camera image in the form of reddish color fringes along edges, which are marked by red arrows in the magnified (and contrast-enhanced) details on the right. The occurrence and strength of such color fringes in general varies with the position of the respective edge in the image (here, they only occur at edges facing away from the image center). The original image of an Apple MacBook Pro keyboard was acquired with a Nikon 18–200 mm zoom lens at a focal length of 150 mm.

only varies across individual sensor elements. This makes spatial noise particularly interesting to forensic investigators as it not only can serve as a 'fingerprint' [137] of a specific digital camera, but can also be tested for consistent appearance across the image [154]. The main components of spatial noise are *photo-response non-uniformity* (PRNU) and *dark current*. PRNU is a multiplicative noise that is caused by differences in the quantum efficiency of individual sensor elements due to inevitable variations in their manufacturing process. Dark current is a thermal effect, which results in free electrons even if the sensor is not exposed to any light. Yet its relative strength depends on the individual sensor elements and, for instance, minute differences in their size.

A standard model of the different noise sources is given in the following equation [96, 32], where function $\mathsf{sensor} : \mathbb{R}^{MN} \to \mathscr{X}^{MN}$ maps the incoming light $\boldsymbol{s}$ to a *raw digital image* $\boldsymbol{r}$:[22]

$$\boldsymbol{r} = \mathsf{sensor}(\boldsymbol{s}) = g^{\gamma}\left(\boldsymbol{\kappa}\,\boldsymbol{s} + \boldsymbol{\eta}\right)^{\gamma} + \boldsymbol{\nu}\,. \tag{2.1}$$

The multiplicative PRNU is represented by vector $\boldsymbol{\kappa}$, whereas $\boldsymbol{\eta}$ subsumes a variety of additive noise terms (including dark current, shot noise and read-out noise), and $\boldsymbol{\nu}$ denotes quantization noise. Scalars g and γ are fixed camera parameters that control gain factor and gamma correction of the output signal, respectively.

Already Heerich [97], in 1998, alluded to the use of sensor noise as a means to identify the specific image acquisition device (fax machines in this specific study). Groundbreaking

22 For simplicity, we assume that raw image and final image share a common alphabet $\mathscr{X}$. In practice, raw images are often stored with a higher bit depth. Moreover, the number of sensor elements usually exceeds the number of image pixels. To speed up demosaicing, $M_e \times N_e$ *effective* pixels, $M_e \geq M$, $N_e \geq N$, are used to generate the final image by circumventing special interpolation rules for border pixels. The number of *recorded* pixels might be even higher when, for instance, a certain number of pixels is cropped to reduce visible vignetting effects in the final image.

investigations into the use of dark current and photo-response non-uniformity by Kurosawa et al. [137, 138] and Fridrich et al. [155, 32], respectively, have since then motivated a large number of follow-up studies on noise-based device identification. Because the sensor output is subject to further in-device processing, which is likely to interfere with subtle noise characteristics (cf. Section 2.1.4), sensor noise would ideally be examined based on the raw sensor output [133]. For example, subsequent color filter interpolation, by integrating over a number of adjacent sensor elements, affects noise characteristics of individual pixels [63, 144]. Empirical results suggest, however, that in particular PRNU is a highly indicative identifying trace in arbitrary digital camera images [89],[23] which may even survive a print-scan cycle [88].

Photo-response non-uniformity has also found applications in the examination of scanned images [84, 122]. Here, the typical line sensor of flatbed scanners repeats spatial noise characteristics along rows. This allows forensic investigators to distinguish between digital camera images and scanned images [23]. We further note that not only the presence of specific noise pattern, but also more general characteristics such as the relative strength [33] and the statistical distribution [93, 124] of sensor noise can serve as identifying traces in forensic settings.

Apart from sensor noise, *sensor defects* are another sensor imperfection of forensic relevance [97, 79]. Sensor defects refer to sensor elements that constantly output too high or too low intensity values. Occurrence and position of these defects are again specific to the individual digital camera and accumulate over time, which also gives rise to applications in temporal forensics [67]. A similar effect is caused by *dust particles* on the sensor protective glass [48, 184]. This type of identifying traces is mainly present in cameras with interchangeable lenses. Yet, both sensor defects and dust particles can be relatively easily corrected for. In general, their appearance also strongly depends on the image content and lens settings, making these device characteristics less suitable for general purpose forensic analysis.

2.6.1.3 Color Filter Array Characteristics

Because the sensor can only measure the intensity of the incoming light, the generation of a color image requires the light to be split up to its corresponding components (typically red, green and blue). Most digital camera designs combine a single sensor with an array of color filters so that different samples of the raw signal represent different color information. Missing color information is then obtained from a demosaicing procedure.

The particular way of how these color filters are arranged is referred to as *CFA configuration*. Different camera models employ different CFA configurations, making this parameter a valuable device characteristic in the forensic analysis of digital images [220, 46, 127]. Although, in theory, the CFA configuration is neither restricted to capture exclusively red (R), green (G) and blue (B) components, nor confined to a particular layout [166], it is often one of the four 2×2 Bayer pattern [10] that is repeated over the entire image plane. Here, two green elements are arranged in a diagonal setup and each one red and one blue element fill up the remaining space, cf. Figure 2.11.

23 Camera manufacturers strive for visually appealing output images and try to reduce sensor noise to a minimum. While a correction for additive dark current by subtraction of a so-called dark frame is relatively straight-forward, the suppression of multiplicative PRNU is more involved and thus typically traded off against processing speed.

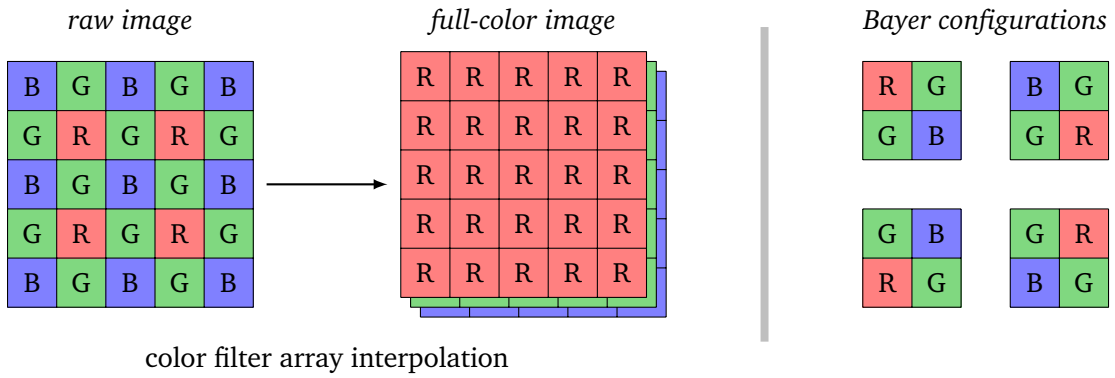


Figure 2.11: Typical digital cameras employ a color filter array (usually one of the four Bayer configurations shown on the right). Each sensor element only records light of a certain wavelength (here red, green and blue). A full-color image is obtained by interpolating the remaining color information from surrounding pixels of the raw image.

Demosaicing implies that only at most one third of all pixels in an RGB image contain genuine information from a sensor element.[24] The remaining pixels are interpolated from neighboring samples of the raw signal,

$$\hat{\boldsymbol{x}} = \texttt{demosaic}\left(\boldsymbol{r}, \boldsymbol{C}\right), \tag{2.2}$$

where function $\texttt{demosaic} : \mathscr{X}^{MN} \times \{\mathrm{R}, \mathrm{G}, \mathrm{B}\}^{MN} \to \mathscr{X}^{3MN}$ is the *demosaicing algorithm*. Matrix $\boldsymbol{C}$ is of dimension $M \times N$ and represents the configuration of the color filter array, and $\hat{\boldsymbol{x}}$ is the demosaiced full-color image. A side-product of CFA interpolation is that neighboring pixels become highly inter-dependent and, as first mentioned by Popescu and Farid [198], the regular structure of typical CFAs leads to periodic variations throughout the entire image. Ho et al. [99] further added that similar dependencies occur between different color channels of the image. Post-processing may wipe out these traces and thus leads to intra-characteristic inconsistencies. Yet the complete absence of CFA artifacts could also be an indication of a non-camera image [75].

The concrete form of demosaicing inter-pixel dependencies depends not only on the CFA configuration, but also on the *demosaicing algorithm*. This observation has motivated Bayram et al. [12] and Swaminathan et al. [220], amongst others, to understand such characteristics as an identifying trace of the specific camera model, or, as noted by Cao and Kot [25], of the raw image processing software used to create the final image.

2.6.1.4 JPEG Compression Characteristics

After CFA interpolation, the demosaiced image undergoes a number of device-internal processing steps, $\boldsymbol{x} = \texttt{process}(\hat{\boldsymbol{x}})$, $\texttt{process} : \mathscr{X}^{3MN} \to \mathscr{X}^{3MN}$. From the forensic investigator's point of view, *JPEG compression* is among the most relevant of these procedures for three reasons. First, JPEG [108] is the quasi-standard for storing digital camera images, as it offers consumers a reasonable trade-off between visual quality and file size. Second, because of its lossy character, JPEG is likely to dilute some of the device characteristics of previous components of the acquisition device. Finally, JPEG compression introduces its very own identifying traces,

24 We say "at most" because sophisticated demosaicing procedures may also 're-compute' samples of the raw signal.

which can be tested both for existence and consistency. These traces are of particular interest because JPEG compression is usually the last step in the in-camera processing chain, i. e., only post-processing outside the camera can impair their existence.[25]

JPEG compression works by dividing the image into non-overlapping blocks $\boldsymbol{x}_{\boxed{i}}$ of 8×8 pixels, where $\boldsymbol{x}_{\boxed{i}}$ denotes the i-th vectorized block. Each block is fed into a two-dimensional discrete cosine transformation (2D-DCT) to yield 64 DCT coefficients, $\boldsymbol{c}_{\boxed{i}} = (c_{\boxed{i}1}, \ldots, c_{\boxed{i}64})$, which are then quantized using coefficient-specific quantization factors, $\dot{q}_j$, $1 \leq j \leq 64$. The strength of compression is controlled by a set of *quantization tables* that are stored with the JPEG image. Farid [55] reports that these tables are a good indicator of the digital camera model. Camera manufacturers are free to fine-tune quantization tables according to their own preferences, and consequently a considerable number of different tables can be found in the wild. Kee et al. [118] further note that no camera model seems to share tables with the widely-used Photoshop image processing toolbox.

By its very definition, quantization in the JPEG compression pipeline affects the *distribution of DCT coefficients*. In particular, quantization introduces gaps to the histogram of DCT coefficients, whereas the width of these gaps depends on the quantization factor, $\dot{q}_j$ (i. e., on the corresponding quantization table). If this 'JPEG fingerprint' [68] is found in an image stored in a lossless format (e. g., PNG or TIFF), the image was very likely JPEG compressed before [52, 157, 156]. The distribution of DCT coefficients is further influenced by processing artifacts due to additional quantization steps (outside the device) [169, 34, 81]. For example, the special case of *double compression* with quantization factors $\dot{q}_j$ and $\ddot{q}_j$ imposes the following relation for the j-th DCT coefficient of the i-th block [196]:

$$\mathring{\mathring{c}}_{\boxed{i}j} = \ddot{q}_j \ddot{c}_{\boxed{i}j} = \ddot{q}_j \left\lfloor \frac{\mathring{c}_{\boxed{i}j}}{\ddot{q}_j} + 1/2 \right\rfloor = \ddot{q}_j \left\lfloor \left\lfloor \frac{c_{\boxed{i}j}}{\dot{q}_j} + 1/2 \right\rfloor \frac{\dot{q}_j}{\ddot{q}_j} + 1/2 \right\rfloor , \tag{2.3}$$

where $\mathring{c}_{\boxed{i}j}$ and $\mathring{\mathring{c}}_{\boxed{i}j}$ denote the dequantized DCT coefficient after the first and second compression step, respectively.[26] Characteristic peaks and gaps occur in the corresponding histogram depending on the specific combination of quantization factors. This can be exploited to detect images that have been re-saved as JPEG multiple times [196, 31, 188, 104, amongst others]. Figure 2.12 illustrates this effect for the particularly indicative case where the second quantization factor is half the first quantization factor, $\dot{q} = 2\ddot{q}$. A typical example of a histogram of double-quantized DCT coefficients is shown in Figure 2.13. Lukáš and Fridrich [152] first noted that such characteristics of histograms of multi-quantized DCT coefficients allow to infer quantization tables of earlier JPEG compression steps (e. g., the table of in-device compression). This has led to a number of related studies [188, 210]. Local inconsistencies due to traces of multi-compression can further serve as indication of localized post-processing [148, 56].

Because each 8×8 block is transformed separately, JPEG compression gives rise to the well-known (and oftentimes visible) *blocking artifacts* in the spatial domain. Similar to DCT coefficient histogram characteristics, such traces can hint to prior JPEG compression of images stored in a lossless format [52, 170]. Blocking artifacts can also be tested for consistency.

25 As more and more digital cameras allow to store images in uncompressed raw format, traces of JPEG compression could likewise be seen as processing artifact. In general, it is up to the forensic investigator to decide whether compression took place inside the device or (possibly after further post-processing steps) outside the device.

26 We ignore rounding and truncation errors that arise in the inverse DCT.

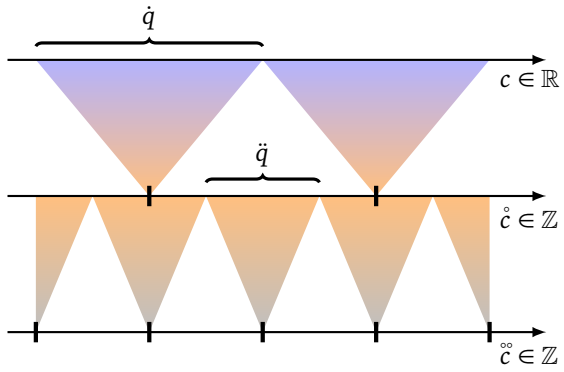


Figure 2.12: Formation of gaps in the DCT coefficient histogram after double-compression with $\dot{q} = 2\ddot{q}$. In the first compression step, real-valued coefficients c are mapped to integers $\mathring{c}$ that are multiples of the corresponding quantization factor $\dot{q}$. In the second compression step, every other multiple of quantization factor $\ddot{q}$ has no single corresponding input value. The shaded triangles visualize coefficient intervals that quantize to the very same discrete value.

Inconsistencies may occur, for example, if parts of a (former) JPEG image are processed or replaced by a non-compressed patch. Moreover, the combination of two or more JPEG images can lead to inconsistent block boundaries when the splicing procedure ignores the regular 8×8 grid and inserts misaligned patches [143].

On a more general level, quantization artifacts are not limited to the specific case of JPEG compression based on 8×8 image blocks. Different source coders may vary in the block size and the employed transformed domain (for instance, Wavelet-based compression in the JPEG2000 standard [109]) and thus introduce their very own identifying traces [146].

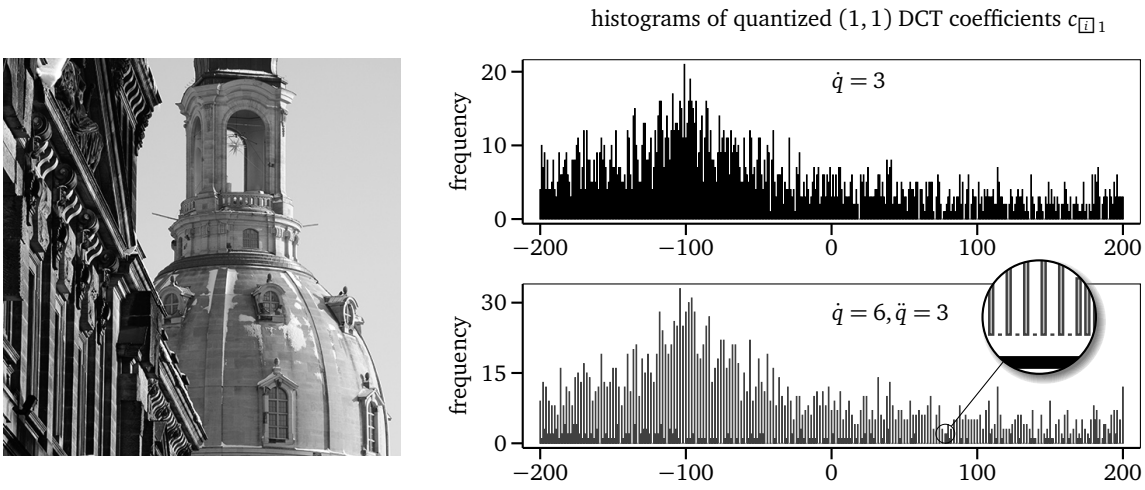


Figure 2.13: Traces of JPEG double compression in the histograms of quantized (1,1) DCT coefficients. Both histograms were obtained from the grayscale image depicted on the left. The upper histogram of coefficients $\dot{c}_{\boxed{i}1}$ corresponds to a single-compressed version stored with JPEG quality 90 ($\dot{q}_1 = 3$). The lower histogram of coefficients $\ddot{c}_{\boxed{i}1}$ belongs to a double-compressed version that was first stored with JPEG quality 80 ($\dot{q}_1 = 6$), followed by a second compression with JPEG quality 90 ($\ddot{q}_1 = 3$). Double-compression introduces characteristic gaps to the DCT coefficient histogram that are not present in the histogram of the single-compressed image. (The original image is part of the Dresden Image Database [82]. To enhance the quality in print only the center part of the histograms is shown.)

2.6.1.5 Camera Response Function

All device characteristics we have discussed so far reflect particularities of specific components of the image acquisition pipeline. However, it is also possible to consider the complete device as a black box (cf. Figure 2.7) and to study its input-output relation, which is also known as *camera response function* or *radiometric response function*, $\mathtt{crf} : \mathbb{R}^{MN} \to \mathscr{X}^{MN}$. This mapping from incoming light to pixel intensities is monotonically increasing and typically of non-linear behavior, the latter of which is also reflected in the gamma correction term, γ, in the sensor noise model in Equation (2.1). Ng et al. [182] used a more flexible parametric model of non-linearity,

$$\mathtt{crf}(\boldsymbol{s}) = \boldsymbol{s}^{\sum_{k=0}^{K} a_k \boldsymbol{s}^k}, \tag{2.4}$$

(yet ignoring noise and quantization effects), which they later extended to include a linear term for low irradiance [178, 174]. As first mentioned by Lin et al. [147] in a forensic context, different camera models are in general expected to have different camera response functions, i. e., they vary in the coefficient vector $\boldsymbol{a}$ of the above model. This gives rise to applications in camera model identification. Image manipulations can be detected by testing for a consistent response function throughout the image, which applies in particular to cases where a number of images from different sources are 'spliced' together [147, 103]. Lin et al. [147] further note that intra-characteristic inconsistencies may also become apparent by examining and comparing the camera response functions of different color channels, which should exhibit a similar shape for original images.

2.6.2 Processing Artifacts

Post-processing can not only be detected by means of inconsistent or missing device characteristics, but also because of specific processing artifacts that certain image processing primitives leave behind in the resulting image. While the former—against the backdrop of requiring knowledge of (or assumptions about) potential acquisition devices—have the advantage of being relatively independent of the actual form of post-processing, the latter allow for the targeted analysis of particular classes of post-processing. This can provide forensic investigators with additional information about the processing history of the image under analysis. Consequently, a number of specific processing artifacts have received interest in the literature, three of which we will discuss in more detail in the following.

2.6.2.1 Copy–Move Manipulations

Among the large class of processing artifacts, traces of *copy–move manipulations* have received particular interest in the literature. A copy–move manipulation copies a part of an image to insert it into another part of the same image, usually with the goal to conceal or emphasize details of the original image. Figure 2.14 gives a typical example and depicts a copy–move manipulation in an image of an Iranian missile test. By its very definition, a copy–move manipulation introduces duplicate regions to the resulting image. Fridrich et al. [69] were the first to understand this as identifying trace in a forensic context. The authors also mentioned that forensic investigators, in a more realistic setting, will rather encounter *near-duplicate* image regions. First, the copied region often needs to be further processed to better align with the surrounding image content. Second, the manipulated image will often be JPEG compressed

Figure 2.14: A typical copy–move manipulation, applied to a press photograph of an Iranian missile test. A non-functioning missile in the presumably original image (shown on the left, source: online service of the Iranian Daily Jamejam Today) was replaced by a copy of another missile (middle, source: Iranian Revolutionary Guards). The manipulated image appeared on the title page of a number of major Western newspapers [183]. The right image shows the output of Popescu and Faird's [195] copy–move detector and marks near-duplicate regions.

prior to publication. This observation has led to a plethora of copy–move detectors, which basically all differ only in their definition of 'near-duplicate'. Specific instances include, for example, the detection of blurred regions [160] or so-called Poisson cloning [49]. The latter yields particularly appealing manipulations by blending the boundaries of the copied region adaptively to its neighborhood [187]. Finally, a more general form of copy–move manipulation may also include a geometric transformation (e. g., scaling or rotation) of the copied region [105, 185, 5]. We refer to two recent surveys by Christlein et al. [39, 40] for a comprehensive overview of relevant approaches to copy–move detection.

2.6.2.2 Resampling

Not only copy–move manipulations often require parts of digital images to be resized or rotated in order to better align with the remaining image content. In general, any *geometric transformation*—whether applied to a specific region of the image or to the image as a whole—is of interest when inferring the processing history of an image under analysis. Technically, a geometric transformation can be described as *resampling* of the original image grid to a new image grid. Interpolation is the key to smooth and visually appealing transformations. However, a virtually unavoidable side-effect of interpolation is that it introduces linear dependencies between groups of adjacent pixels. Popescu and Farid [197] and Gallagher [74] first realized that these dependencies periodically vary throughout the image and thus can be understood as identifying trace of a previous resampling operation. The period length in general depends on the transformation parameters, which allows for inference about the specific transformation applied to the image [126]. Figure 2.15 illustrates the formation of periodic linear dependencies for the particularly indicative case of upscaling by a factor of two using bilinear interpolation. It was later noted [193, 236, 132] later noted that resampling may also interfere with characteristics of previous JPEG compression. More specifically, a geometric transformation not only maps the image itself to a new image grid, but also affects the shape of existing JPEG blocking artifacts. In the above example of upsampling by a factor of two, transformed "JPEG blocks" will have a size of 16×16 instead of 8×8 pixels, which can be interpreted as a further indication of resampling.[27]

27 In this specific example—and in general for any integer scaling factor—a subsequent JPEG compression will "restore" the original 8×8 grid.

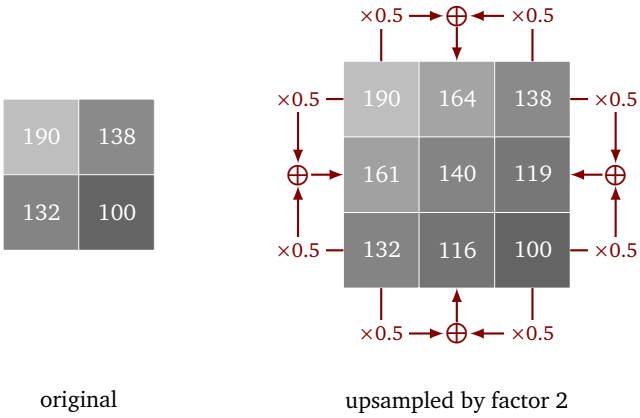


Figure 2.15: Upsampling of a 2×2 image block (left) by a factor of two using bilinear interpolation (right). Pixels in every other row (or column) are linear combinations of their direct vertical (or horizontal) neighbors. Because a resized image is composed of a large number of such blocks, periodic linear dependencies occur.

2.6.2.3 Contrast Enhancement

Our last example of specific processing artifacts concerns identifying traces of the general class of pixel intensity mappings

$$x_i \mapsto \lfloor \mathrm{f}(x_i) + 1/2 \rfloor \,, \tag{2.5}$$

where $\mathrm{f} : \mathscr{X} \to \mathscr{X}$ is a monotone and non-linear function. Such mappings (with $\mathrm{f}(\cdot)$ being monotonically increasing) are commonly used for *contrast enhancement* of digital images and are thus an important image processing primitive. Gamma correction is a special case, where function $\mathrm{f}(\cdot)$ takes the form

$$\mathrm{f}(x_i) = (2^\ell - 1) \left(\frac{x_i}{2^\ell - 1} \right)^\gamma . \tag{2.6}$$

Although contrast enhancement alone typically will not impair the authenticity of an image, the detection thereof is still of high relevance in digital image forensics for several reasons (apart from the general question about the processing history of the image under analysis). For example, many forensic techniques rely on some form of linearity assumption (e. g., the assumption of linear inter-pixel dependencies in resampling detection, cf. Section 2.6.2.2 above), which may not hold after non-linear processing. Moreover, contrast enhancement may be part of a more complex manipulation, where it is applied locally with the objective to adjust manipulated regions to their surrounding.

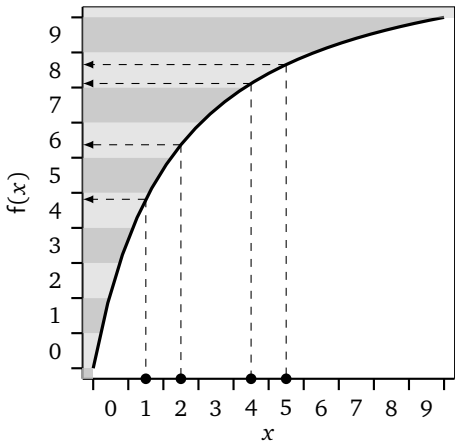


Figure 2.16: Formation of peaks and gaps in the histogram of contrast-enhanced digital images. A peak occurs if multiple discrete input values, after rounding, map to the same output value (here, for example, $\mathrm{f}(4) = \mathrm{f}(5) = 8$). On the contrary, a gap occurs if a specific output value has no single corresponding input value (here, for instance, $\mathrm{f}(x) = 5$). The alternating dark-gray and light-gray regions correspond to rounding intervals that map to the same output value.

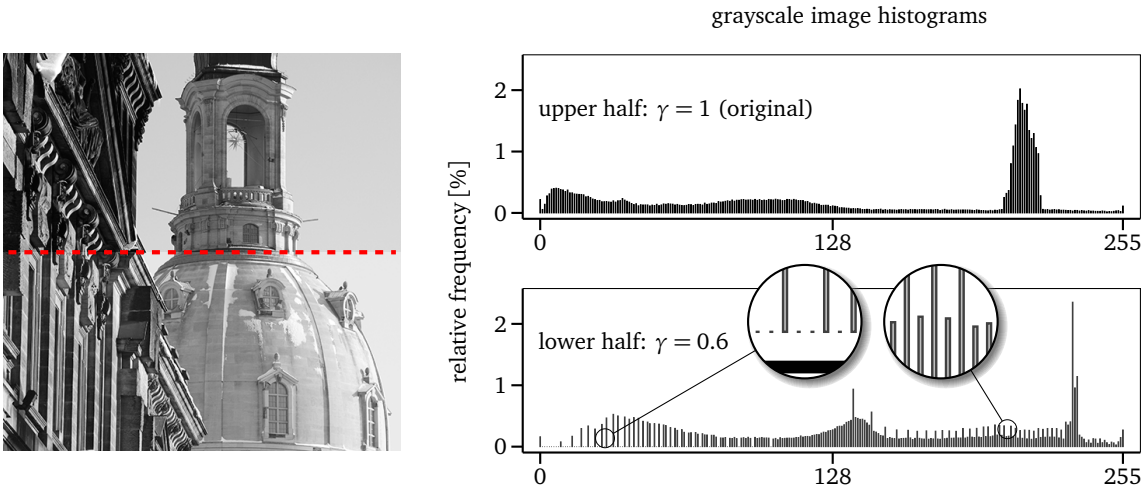


Figure 2.17: Typical grayscale histograms before and after gamma correction. The upper half of the image was left unmodified, whereas the lower half was subject to contrast reduction with $\gamma = 0.6$. Gamma correction introduces characteristic peaks and gaps to the histogram that are not present in the histogram of the upper part of the image. (The original image is part of the Dresden Image Database [82].)

Because contrast enhancement modifies intensity values of individual pixels, it is not without influence on the histogram of the resulting image. More specifically, Stamm and Liu [211] observed that histograms of contrast-enhanced images exhibit characteristic peaks and gaps, which are not present in histograms of typical original digital image. Figure 2.16 visualizes how the combination of non-linearity and rounding in Equation (2.5) leads to these identifying traces. The accompanying Figure 2.17—for the special case of gamma correction—compares the histogram of an original image to the histogram of a contrast-enhanced image. Stamm and Liu [212] further note that the position of peaks and gaps in the histogram depends on the specific form of contrast enhancement, which allows to infer parameters of the mapping $\mathsf{f}(x)$.

3 A Unified Digital Image Forensics Framework

This chapter devises a formal framework of digital image forensics. Its objective is to provide a consistent terminology to foster a discussion of the general properties of image forensics algorithms and counter-forensic techniques. To the best of our knowledge, the only similar attempt to formalize theoretical underpinnings of certain aspects of digital image forensics was made in a series of papers by Swaminathan et al. [218, 219, 220, 221, 222]. The authors extensively studied the problem of *component forensics*, which they defined as the approach to identify the *«algorithms and parameters employed in the various components of the device that was used in capturing the data»* [224]. Here, we take a broader approach and do not limit ourselves to the imaging device, but rather consider the complete image generation process (see Section 3.5.2 for a discussion of relations to the work of Swaminathan et al.).

3.1 Image Generation Formalized

Digital image forensics exploits inherent characteristics of the image generation process and their impact on the resulting image under analysis. For a formal treatment of image forensics we thus have to define and characterize this process and its inputs (Section 3.1.1) before we can reason about the notion of image authenticity (Section 3.1.2).

3.1.1 Image Generation Process

Digital images are projections of observations of the infinite set of all conceivable scenes $\mathcal{S} \in \mathscr{S}$ to vectors $\boldsymbol{x} \in \mathbb{X}$, $\mathbb{X} \equiv \mathscr{X}^N$, over a finite alphabet $\mathscr{X}$ of discrete symbols. An universal *image generation function* helps us to conveniently formalize such projections.

Definition 1 | Image Generation Function. Function $\texttt{generate} : \mathscr{S} \times \Theta \to \mathbb{X}$ maps real-world phenomena $\mathcal{S} \in \mathscr{S}$ to digital images $\boldsymbol{x} \in \mathbb{X}$. The mapping is parametrized with a collection of parameters $\boldsymbol{\theta} \in \Theta$.

The parameters include, inter alia, the perspective and the time of acquisition. They also reflect different generation process characteristics, such as the choice of a particular acquisition device and its configuration (e. g., settings, lenses) and control how the image is processed after acquisition. In its simplest form, it is convenient to think of this image generation function as a combination of both, initial image acquisition and subsequent post-processing.

Definition 2 | Image Acquisition and Image Processing. Function `generate` is a concatenation of an *image acquisition function* $\texttt{acquire} \in \mathscr{A} : \mathscr{S} \to \mathbb{X}$ and an *image processing function* $\texttt{process} \in \mathscr{P} : \mathbb{X}^+ \to \mathbb{X}$, where, for a set $\mathscr{P}_0$ of elementary image processing primitives, $\mathscr{P}$ is given by $\mathscr{P} = \mathscr{P}_0^+$. Tuples $(\texttt{acquire}, \texttt{process})$ are elements of the function space $\mathscr{A} \times \mathscr{P}$ of combinations of all possible image acquisition methods and all possible image processing operations, respectively. Their exact composition is defined by the parameters $\boldsymbol{\theta}$ of `generate`.

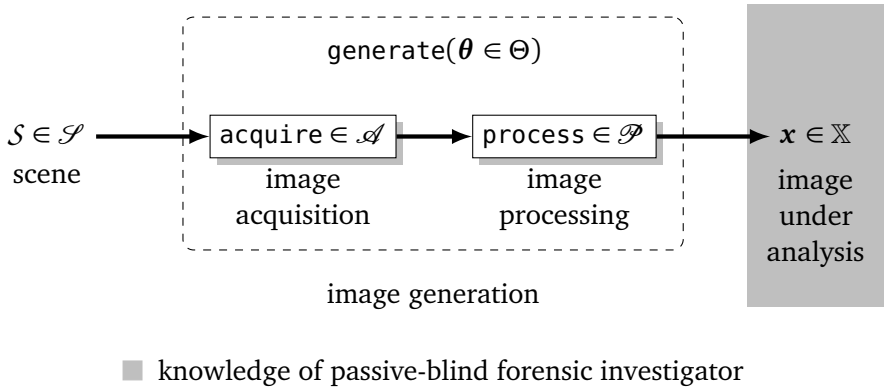


Figure 3.1: Universal image generation process in the context of passive–blind image forensics.

In the above definition, operator $^+$ is the 'Kleene plus', which for a given set $\mathbb{X}$ is defined as $\mathbb{X}^+ = \bigcup_{n=1}^{\infty} \mathbb{X}^n$. Hence, function `process` may take an arbitrary positive number of digital images as input. Set $\mathscr{P}_0 \subset \mathscr{P}$ forms the basis for the construction of arbitrary image processing operations. It can be interpreted as a collection of basic components, such as linear filtering or compression. These image processing primitives are atomic in the sense that they cannot be split up further into a sequence of independent sub-processes. Using this notation, a tuple $\texttt{process} = (\texttt{process}_1, \ldots, \texttt{process}_n) \in \mathscr{P}_0^n$ means that, after initial acquisition, a series of n independent processing steps leads to the final image $\boldsymbol{x} \in \mathbb{X}$.

The block diagram in Figure 3.1 illustrates our idea of an universal image generation process (ignoring the possibility of multiple inputs to function `process` for the sake of simplicity), which serves as a basis for the following considerations. Intuitively, function `acquire` directly relates to the digital imaging device used to initially capture the image. Function `process`, on the other hand, refers to all post-processing outside the device. However, we will see in Section 3.3.3.3 that practical settings may require more flexible definitions whenever device-internal processing occurs.

3.1.1.1 Special Case: Digital-to-Analog-to-Digital Conversion

To fit function `generate` into the notion of a more general image generation process introduced informally in Figure 2.1, we also have to consider images that undergo a digital-to-analog transformation, followed by a re-acquisition with a (possibly different) digital imaging device.

Definition 3 | Analog Output. Functions $\texttt{output} \in \mathscr{O} : \mathbb{X} \to \mathscr{S}_{\mathscr{O}}$ map digital images to natural phenomena. Set $\mathscr{S}_{\mathscr{O}} \subset \mathscr{S}$ contains all conceivable natural phenomena that can result from digital-to-analog conversions. Hence, tuples $(\texttt{output}, \texttt{acquire})$ are elements of the space $\mathscr{O} \times \mathscr{A} \subset \mathscr{P}_0$ of all possible *digital-to-analog-to-digital conversions*.

Observe that we consider the space $\mathscr{O} \times \mathscr{A}$ as a subset of $\mathscr{P}_0$. This is in accordance to our comment in Section 2.1.1 that re-digitization can be seen as a special form of post-processing. Because function `process` now "hides" the possibly infinite digital-to-analog-to-digital loop of Figure 2.1, instances of `process` can encompass sub-procedures that are itself an image generation function according to Definitions 1 and 2.

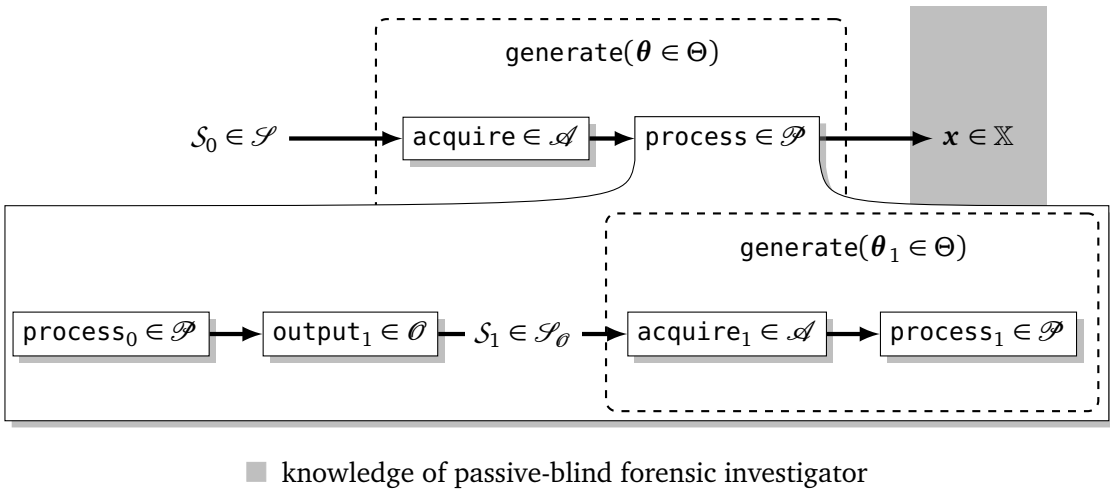


Figure 3.2: Digital image generation with digital-to-analog-to-digital conversion. The general image processing function process is a concatenation of several sub-procedures, including a nested instance of generate.

Figure 3.2 gives a typical example, where the captured image is first fed into a function $\texttt{process}_0 \in \mathscr{P}$ before being transformed to the analog domain via function $\texttt{output}_1 \in \mathscr{O}$. The so-generated natural phenomenon $\mathcal{S}_1 \in \mathscr{S}_{\mathscr{O}}$ is then re-digitized by function $\texttt{acquire}_1 \in \mathscr{A}$, which may be succeeded by a further processing step $\texttt{process}_1 \in \mathscr{P}$, leading to the final image $\boldsymbol{x} \in \mathbb{X}$. Tuple $(\texttt{acquire}_1, \texttt{process}_1)$ encapsulates a *nested image generation function*. Its parameters $\boldsymbol{\theta}_1 \in \Theta$ can be derived from the parameter set of the 'parent' process.

Note that the above framework generalizes to any desirable number of nested image generation functions: both functions $\texttt{process}_0$ and $\texttt{process}_1$ can be further specified to include arbitrary sub-processes. The forensic investigator, however, has generally only access to the result of the overall process.

3.1.1.2 Special Instances of generate

There exist two parameter settings that deserve a special note. First, digital images not necessarily undergo a processing step after initial acquisition with a digital imaging device. Set $\mathscr{P}_0$ thus explicitly includes the identity function $\perp_{\mathscr{P}} : \boldsymbol{x} \mapsto \boldsymbol{x}$, i. e., no post-processing.

Definition 4 | Original and Processed Images. All non-nested digital image generation functions[28] $(\texttt{acquire}, \perp_{\mathscr{P}}) \in \mathscr{A} \times \mathscr{P}$ produce *original images* as opposed to *processed images* that result from generation functions $(\texttt{acquire}, \texttt{process}) \in \mathscr{A} \times \mathscr{P} \setminus \{\perp_{\mathscr{P}}\}$.

Similarly, set $\mathscr{A}$ includes a pathologic function $\perp_{\mathscr{A}}$, which refers to no acquisition with an imaging device. This is particularly useful to differentiate between natural images and computer-generated images.

28 The restriction to non-nested image generation functions is needed to rule out image modifications due to digital-to-analog-to-digital conversions and optional prior processing, cf. Figure 3.2. In practice, this rather strict definition is often replaced by one that accepts any acquired image as original.

Definition 5 | Natural and Computer-Generated Images. All digital image generation functions $(\texttt{acquire}, \texttt{process}) \in \mathscr{A} \setminus \{\perp_{\mathscr{A}}\} \times \mathscr{P}$ generate *natural images* as opposed to *computer generated images* that result from generation functions $(\perp_{\mathscr{A}}, \texttt{process}) \in \mathscr{A} \times \mathscr{P}$. By definition, computer-generated images are processed images.

Note that other research fields define natural images as images of 'natural scenes' and attempt to narrow down characteristics of such phenomena and digital projections thereof [107, i. a.]. We refrain from unnecessarily restricting our framework to projections of particular types of scenes (and the question of what is considered 'natural'). Instead, we focus on particulars of the image generation process.

3.1.2 Authenticity and Semantic Meaning of Digital Images

Two important attributes with regard to the image generation process are the notions of authenticity and semantic meaning of digital images. *Authentic* here means that an image $\boldsymbol{x}$ is a valid projection of the natural phenomenon $\mathcal{S}$, whereas instances of `process` may impair its authenticity. The question whether an image is authentic is deeply entangled with its *semantic meaning*, which refers to the relationship between a depicted scene and the corresponding natural phenomenon.

Before we give a formal definition of authenticity, we note that, intuitively, the projection of one particular natural phenomenon $\mathcal{S}$ to an authentic image is not necessarily unique. More specifically, there may exist a whole set of mappings that yield *semantically equivalent* images. This means that each element in a set of semantically equivalent images $\boldsymbol{x}_1 \neq \boldsymbol{x}_2 \neq \cdots \neq \boldsymbol{x}_n$ is a valid representation of the same realization of $\mathcal{S}$, i. e., it shares the very same scene characteristics with all other images of this set (cf. Section 2.1.4). As a typical example, consider the case where the same scene is captured with many different digital cameras. While each camera will give a slightly different result (depending on specific generation process characteristics), all these images share the same scene characteristics and may be regarded as authentic. Within certain limits also the change of resolution or lossy compression may retain an image's authenticity. In this sense, authenticity is an attribute of tuples $(\boldsymbol{x}, \boldsymbol{\theta}, \mathcal{S})$ where $\mathcal{S}$ must be the realization of $\mathscr{S}$ under parameters $\boldsymbol{\theta}$.

Because the association of semantic meaning requires (human) interpretation and is highly dependent on the context in general, it is exceedingly difficult, if not impossible, to find a closed formalization. Here, we work around this difficulty and assume that semantic equivalence is measurable between images.

Definition 6 | Semantic Equivalence. Two images $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{X}$ are semantically equivalent if there exists a scene $\mathcal{S} \in \mathscr{S}$ such that

$$\left|\mathsf{semantic.dist}(\boldsymbol{x}_1, \mathcal{S}) - \mathsf{semantic.dist}(\boldsymbol{x}_2, \mathcal{S})\right| < d, \tag{3.1}$$

where $\mathsf{semantic.dist} : \mathbb{X} \times \mathscr{S} \to \mathbb{R}_+$ is a measure of the *semantic distance* between image $\boldsymbol{x}$ and a scene $\mathcal{S}$, and d is a given threshold.

The *semantic resolution* is a measure of the ability of function $\mathsf{semantic.dist}$ to differentiate between very similar real-world phenomena for a fixed image $\boldsymbol{x}$. It depends on the *quality* of an image, or, more precisely, on the information conveyed in an image $\boldsymbol{x}$ about $\mathcal{S}$. Threshold d

$\boldsymbol{x}_1$ $\mathring{\boldsymbol{x}}_1$ $\mathring{\boldsymbol{x}}_2$ $\boldsymbol{x}_2$

Figure 3.3: High-quality images $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, together with respective low-quality quantized versions $\mathring{\boldsymbol{x}}_1$ and $\mathring{\boldsymbol{x}}_2$. Higher quality results in higher semantic resolution and vice versa. The original image is part of the Dresden Image Database [82].

in the above definition has to be chosen commensurate with the semantic resolution of the image with the lowest quality.

Figure 3.3 gives a practical example by depicting high-resolution images $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, accompanied by respective low-resolution versions $\mathring{\boldsymbol{x}}_1$ and $\mathring{\boldsymbol{x}}_2$. A high visual quality of the first pair of images makes it relatively save to conjecture that they are not semantically equivalent. As to the pair of low-resolution images, strong quantization increases uncertainty and consequently, a clear distinction with respect to the originating natural phenomenon cannot be drawn. Also, with the quality of images $\mathring{\boldsymbol{x}}_1$ and $\mathring{\boldsymbol{x}}_2$ being inferior to that of image $\boldsymbol{x}_1$, each of the corresponding image pairs $(\boldsymbol{x}_1, \mathring{\boldsymbol{x}}_1)$ and $(\boldsymbol{x}_1, \mathring{\boldsymbol{x}}_2)$ could be likewise considered semantically equivalent.

Equipped with the notion of semantic equivalence, we can finally define what qualifies an image as authentic, i. e., which projections $\boldsymbol{x}$ are valid representations of reality.

Definition 7 | Authentic Images. All original natural images are authentic. Furthermore, for a given authentic image $\boldsymbol{x}_1 = \texttt{generate}(\mathcal{S}, \boldsymbol{\theta})$, a processed version $\boldsymbol{x}_2 = \texttt{process}(\boldsymbol{x}_1)$ is called authentic if $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are semantically equivalent with respect to $\mathcal{S}$.

Definitions 4 and 7 reflect the subtle yet significant difference between processed and inauthentic images. While each non-trivial instance of `process` damages the originality of an image, it not necessarily impairs its authenticity. Whether or not a processed image will be considered inauthentic ultimately depends on a given context and established habits, cf. Section 3.3.3. We further point out that computer-generated images are not inauthentic by definition, because function `process` can always be defined to replace a natural image with a computer-generated version (or parts thereof). This is viable as long as synthesis algorithms are sophisticated enough to generate semantically equivalent images. Similarly, images that underwent a digital-to-analog-to-digital conversion may retain their authenticity under certain circumstances.

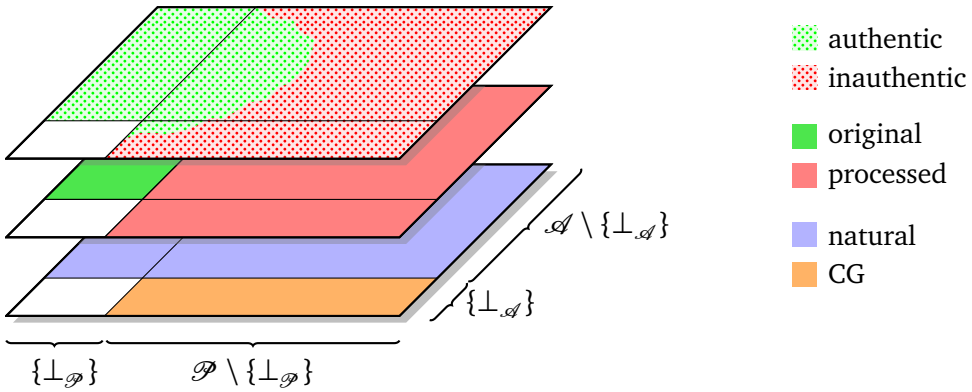


Figure 3.4: Function space $\mathscr{A} \times \mathscr{P}$ of conceivable combinations of image acquisition and processing functions along with different classifications of the resulting digital images.

Our remarks are also reflected in Figure 3.4, which gives a schematic overview of the different classifications of digital images that we encountered previously. Each of the three layers depicts different partitions of the same function space $\mathscr{A} \times \mathscr{P}$, whereas four rectangular regions mark the possible combinations of pathologic and non-trivial acquisition and processing functions, respectively, cf. Definitions 4 and 5. The square in the lower left corner is left blank intentionally, as processing functions $(\perp_{\mathscr{A}}, \perp_{\mathscr{P}})$ have no practical and meaningful equivalent.

3.2 Digital Image Forensics as a Classification Problem

Identifying traces of image generation processes reflect variations in the parameter settings $\boldsymbol{\theta}$ of `generate`, and specific characteristics that are common to images generated with a certain subset $\Theta^{(k)} \subset \Theta$ of parameters. It is straightforward to model these variations in terms of different *classes* of generation processes (Section 3.2.1), and we already encountered some intuitive classifications in the previous section, cf. Figure 3.4. Forensic investigators then strive to make a *decision* on the correct class an image under analysis (via its generation process) belongs to (Section 3.2.2). Hence, digital image forensics is best described as a classification problem.

3.2.1 Classes in Digital Image Forensics

Forensic investigators define classes to encapsulate parameter ranges of the function `generate`. Often, the partition of the function space according to one of the three layers in Figure 3.4 will be too coarse, and the choice of the class space, denoted by $\mathscr{C} = \{\mathcal{C}_0, \ldots, \mathcal{C}_{K-1}\}$, rather depends on the concrete application. For example, manipulation detection is usually stated as binary classification problem, i. e., $|\mathscr{C}| = 2$, with one class $\mathcal{C}_1$ for authentic images and another class $\mathcal{C}_0$ for inauthentic images. For source identification however, different classes represent different imaging sources, e. g., specific device models or individual devices (typically $|\mathscr{C}| \gg 2$).

Definition 8 | Class. A class $\mathcal{C} \in \mathscr{C}$ partitions the function space $\mathscr{A} \times \mathscr{P}$ into two subspaces, $(\mathscr{A} \times \mathscr{P})^{(\mathcal{C})}$ and $(\mathscr{A} \times \mathscr{P})^{(\mathcal{Q})}$, so that all images $\boldsymbol{x}^{(\mathcal{C})}$ generated by $(\texttt{acquire}, \texttt{process}) \in (\mathscr{A} \times \mathscr{P})^{(\mathcal{C})}$ share common identifying traces, $\boldsymbol{x}^{(\mathcal{C})} \in \left\{ \boldsymbol{x} \,\middle|\, \boldsymbol{x} = \texttt{generate}\left(\boldsymbol{\theta} \in \Theta^{(\mathcal{C})}\right) \right\}$.

Convention. To keep notations simple, we use $\boldsymbol{x}^{(k)}$ equivalent for $\boldsymbol{x}^{(\mathcal{C}_k)}$ when referring to images of a particular class $\mathcal{C}_k \in \mathscr{C}$. Moreover, we write $\boldsymbol{x}^{(1)}$ for authentic images and $\boldsymbol{x}^{(0)}$ for inauthentic images whenever the context prevents ambiguities and the class space contains only these two classes.

Definitions 1–8 allow us to conveniently express various problems studied in the course of digital image forensics in a unified formal framework (also see Section 2.4). The most relevant problems are illustrated by the following examples.

Example 2 | Natural versus Computer-Generated Images. Class $\mathcal{C}_1$ of natural images contains all instances of images $\boldsymbol{x}^{(1)}$ generated by functions in the subspace $\mathscr{A} \setminus \{\perp_{\mathscr{A}}\} \times \mathscr{P}$. Class $\mathcal{C}_0$ of computer-generated images entails all instances of images $\boldsymbol{x}^{(0)}$ generated by functions in the subspace $\{\perp_{\mathscr{A}}\} \times \mathscr{P}$.

Example 3 | Source Device Identification. Class $\mathcal{C}_k$ corresponds to acquisition with device k and contains all instances of images $\boldsymbol{x}^{(k)}$ generated by functions in the subspace $\mathscr{A}_k \times \mathscr{P}$. Set $\mathscr{A}_k \subset \mathscr{A}$ entails all image acquisition functions of device k, where $\bigcap_k \mathscr{A}_k = \emptyset$.

The above example can be generalized to arbitrary levels of identification granularity to handle device, model, or type identification, cf. Section 2.4.2.1.

Example 4 | Detection of Processing Artifacts. Class $\mathcal{C}_1$ of original images contains all instances of images $\boldsymbol{x}^{(1)}$ generated by functions in the subspace $\mathscr{A} \times \{\perp_{\mathscr{P}}\}$. Class $\mathcal{C}_0$ of processed images entails all instances of images generated by functions in the subspace $\mathscr{A} \times \mathscr{P} \setminus \{\perp_{\mathscr{P}}\}$. Specific instances of this problem exploit missing or inconsistent device characteristics of acquisition functions $\texttt{acquire} \in \mathscr{A}_k \subset \mathscr{A}$. This effectively restricts the analysis to subspaces $\mathscr{A}_k \times \mathscr{P}$.

Example 5 | Printer Forensics. Class $\mathcal{C}_k$ corresponds to images that have been printed with printer k and contains all instances of images $\boldsymbol{x}^{(k)}$ generated by functions in the subspace $\mathscr{A} \times \big(\mathscr{P} \times (\mathscr{O}_k \times \mathscr{A})\big)$. Set $(\mathscr{O}_k \times \mathscr{A}) \subset \mathscr{P}_0$ refers to all digital-to-analog-to-digital conversions with printer k.

The more general problem to detect re-captured images is a special case of Example 4. Here, class $\mathcal{C}_1$ corresponds to generation functions in the subspace $(\mathscr{O} \times \mathscr{A}) \subset \mathscr{P}_0$.

Example 6 | Temporal Forensics. Class $\mathcal{C}_{t_1,t_2}$ relates to images acquired in the time interval $t_1 < t < t_2$. It contains all instances of images $\boldsymbol{x}^{(\mathcal{C}_{t_1,t_2})}$ generated by functions in the subspace $\mathscr{A}_{t_1,t_2} \times \mathscr{P}$ where $\mathscr{A}_{t_1,t_2} \subset \mathscr{A}$ is the set of all image acquisition functions invoked between time t_1 and t_2.

Note that classes are intentionally defined to partition the parameter space of the image generation process, not the image space. This is why the above examples refer to *instances* of images, i. e., outputs of specific invocations of `generate`. As a result, a given image $\boldsymbol{x} \in \mathbb{X}$ with unknown provenance may be the outcome of different generation functions spanning more than one class. Such ambiguities in this possibilistic framework can be resolved by using a probabilistic perspective. More specifically, we interpret each class $\mathcal{C}$ to define a probability space $(\Omega, \mathcal{P}_{\mathcal{C}})$, with $\Omega = \mathbb{X}$ and $\mathcal{P}_{\mathcal{C}}$ being the *class likelihood*.

Definition 9 | Class Likelihood. Function $\mathcal{P}_{\mathcal{C}} : 2^{\mathbb{X}} \to [0,1]$ is the likelihood function that returns the conditional probability $\mathcal{P}_{\mathcal{C}}(\mathbb{X}_0) = \Pr(\mathbb{X}_0 \,|\, \mathcal{C})$ of observing a subset of images $\mathbb{X}_0 \subseteq \mathbb{X}$ if their generation processes fall in the partition of class $\mathcal{C}$. The probability depends on the empirical distributions $\mathcal{S} \sim \mathscr{S}$ and $(\texttt{acquire}, \texttt{process}) \sim (\mathscr{A} \times \mathscr{P})^{(\mathcal{C})}$.

Convention. To simplify notation, we use $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}) = \Pr(\boldsymbol{x} \,|\, \mathcal{C})$ equivalent for $\mathcal{P}_{\mathcal{C}}(\{\boldsymbol{x}\})$.

3.2.2 Decision Rules

Given a class space $\mathscr{C}$, $K = |\mathscr{C}| \geq 2$, and an observed image $\boldsymbol{x}$ of unknown provenenance, the forensic investigator strives to assign $\boldsymbol{x}$ to a class $\mathcal{C}_*$ according to a decision rule, which is defined to make the best possible decision with respect to some optimality criterion.

Definition 10 | Digital Image Forensics Algorithm. A *digital image forensics algorithm* is a decision rule $\texttt{decide} : \mathbb{X} \to \mathscr{C}$ that assigns an image $\boldsymbol{x} \in \mathbb{X}$ to a class $\mathcal{C}_* \in \mathscr{C}$.

Function `decide` now partitions the image space into disjoint regions $\mathbb{X} = \bigcup_k \mathscr{R}_k$, with $\mathscr{R}_k \cap \mathscr{R}_l = \emptyset$ for all $k \neq l$, such that all elements within a *decision region* $\mathscr{R}_k$ are assigned to class $\mathcal{C}_k$,

$$\mathscr{R}_k = \{\boldsymbol{x} \in \mathbb{X} \,|\, \texttt{decide}(\boldsymbol{x}) = \mathcal{C}_k\}\,. \tag{3.2}$$

It is reasonable to assume that decisions are based on the the class probabilities conditional to the observed image, $\Pr(\mathcal{C}_k \,|\, \boldsymbol{x})$, which reflect the probability that $\boldsymbol{x}$ exhibits identifying traces of generation functions $(\texttt{acquire}, \texttt{process}) \in (\mathscr{A} \times \mathscr{P})^{(\mathcal{C}_k)}$. According to Bayes' theorem, these *a posteriori probabilities* can be obtained from the class likelihoods in Definition 9 as follows,

$$\Pr(\mathcal{C}_k \,|\, \boldsymbol{x}) = \frac{\Pr(\boldsymbol{x} \,|\, \mathcal{C}_k) \cdot \Pr(\mathcal{C}_k)}{\sum_i \Pr(\boldsymbol{x} \,|\, \mathcal{C}_i) \cdot \Pr(\mathcal{C}_i)} = \frac{\mathcal{P}_{\mathcal{C}_k}(\boldsymbol{x}) \cdot \Pr(\mathcal{C}_k)}{\sum_i \mathcal{P}_{\mathcal{C}_i}(\boldsymbol{x}) \cdot \Pr(\mathcal{C}_i)}\,, \tag{3.3}$$

$\Pr(\mathcal{C}_k)$ denotes the corresponding class *prior probabilities*, whereas $\sum_k \Pr(\mathcal{C}_k) = 1$. In general, the larger is $\Pr(\mathcal{C}_k \,|\, \boldsymbol{x})$, the more evidence exists that $\boldsymbol{x}$ was generated by a function $(\texttt{acquire}, \texttt{process}) \in (\mathscr{A} \times \mathscr{P})^{(\mathcal{C}_k)}$. The concrete transformation of posterior probabilities into decisions depends on the algorithm `decide` and its decision rule.

3.2.2.1 Two-Class Decisions

We have already seen in the above examples in Section 3.2.1 that the class space in many forensic problems is defined to comprise only two classes, $\mathcal{C}_0$ and $\mathcal{C}_1$. In fact, also most multi-class forensic analysis are framed as (a series of) binary classification problem(s) by testing each class against all others ("all against all"), or by testing one particular class against a joint class that encompasses all remaining outcomes ("one against all").

Tests for the presence of identifying traces of *one* particular class (assume $\mathcal{C}_k$) will decide $\mathcal{C}_* = \mathcal{C}_{|k-1|}$ if $\boldsymbol{x}$ does not evince (enough of) these characteristics, $k \in \{0,1\}$. A *misclassification* occurs whenever an image $\boldsymbol{x}^{(k)}$ is assigned to the wrong class $\mathcal{C}_{|k-1|}$, and forensic investigators generally wish to operate algorithms `decide` that reduce the number of false decisions to a minimum. In a two-class setting, we may encounter two different types of error. A *miss* refers to an image $\boldsymbol{x}^{(k)}$ that is element of decision region $\mathscr{R}_{|k-1|}$, whereas a *false alarm* occurs when

Table 3.1: Possible outcomes and error probabilities of digital image forensics algorithms defined on a binary class space $\mathscr{C} = \{\mathcal{C}_0, \mathcal{C}_1\}$, when testing for identifying traces of class $\mathcal{C}_k$, $k \in \{0,1\}$.

	examination of images ...	
decision region	$\boldsymbol{x}^{(k)}$	$\boldsymbol{x}^{(\lvert k-1\rvert)}$
$\boldsymbol{x} \in \mathscr{R}_k$	correct detection $1 - P_M$	false alarm P_{FA}
$\boldsymbol{x} \in \mathscr{R}_{\lvert k-1\rvert}$	miss P_M	correct rejection $1 - P_{FA}$

an image $\boldsymbol{x}^{(|k-1|)}$ is contained in decision region $\mathscr{R}_k$, see also Table 3.1. The overall *probability of error* is hence given by

$$\begin{aligned} P_e &= \Pr\left(\mathcal{C}_k\right) \cdot \sum_{\boldsymbol{x} \in \mathscr{R}_{|k-1|}} \mathcal{P}_{\mathcal{C}_k}(\boldsymbol{x}) \;+\; \Pr\left(\mathcal{C}_{|k-1|}\right) \cdot \sum_{\boldsymbol{x} \in \mathscr{R}_k} \mathcal{P}_{\mathcal{C}_{|k-1|}}(\boldsymbol{x}) \\ &= \underbrace{\Pr\left(\mathcal{C}_k\right) \cdot \sum_{\boldsymbol{x} \in \mathbb{X}} \mathcal{P}_{\mathcal{C}_k}(\boldsymbol{x}) \cdot \delta_{\mathtt{decide}(\boldsymbol{x}),\mathcal{C}_{|k-1|}}}_{P_M} + \underbrace{\Pr\left(\mathcal{C}_{|k-1|}\right) \cdot \sum_{\boldsymbol{x} \in \mathbb{X}} \mathcal{P}_{\mathcal{C}_{|k-1|}}(\boldsymbol{x}) \cdot \delta_{\mathtt{decide}(\boldsymbol{x}),\mathcal{C}_k}}_{P_{FA}}, \end{aligned} \tag{3.4}$$

where P_M and P_{FA} denote probability of miss and false alarm, respectively.

Clearly, a forensic investigator's goal is now to find partitions $\mathscr{R}_k$ that minimize the probability of error, or some related measure. Because any decision ultimately depends on the probabilities $\Pr\left(\mathcal{C}_0 \,|\, \boldsymbol{x}\right)$ and $\Pr\left(\mathcal{C}_1 \,|\, \boldsymbol{x}\right) = 1 - \Pr\left(\mathcal{C}_0 \,|\, \boldsymbol{x}\right)$, it is straightforward to specify algorithms `decide` via a threshold parameter $0 \leq \tau' \leq 1$:

$$\mathtt{decide}(\boldsymbol{x}) = \begin{cases} \mathcal{C}_k & \text{for } \Pr\left(\mathcal{C}_k \,|\, \boldsymbol{x}\right) \geq \tau' \\ \mathcal{C}_{|k-1|} & \text{else.} \end{cases} \tag{3.5}$$

Forensic investigators may impose additional constraints to ensure a reliable decision, for example by requiring a minimum separability P_{sep} between the classes,

$$2 \cdot \Pr(\mathcal{C}_* \,|\, \boldsymbol{x}) - 1 \geq P_{\text{sep}}\,. \tag{3.6}$$

In classification theory, this is also known as the *reject option* [38, 14], and function `decide` needs to be refined accordingly to return a special value for undecidable cases.

Because $\Pr\left(\mathcal{C}_0 \,|\, \boldsymbol{x}\right) + \Pr\left(\mathcal{C}_1 \,|\, \boldsymbol{x}\right) = 1$, the decision criterion in Equation (3.5) equates to

$$\Pr\left(\mathcal{C}_k \,|\, \boldsymbol{x}\right) \geq \tau' \quad \Leftrightarrow \quad \frac{\Pr\left(\mathcal{C}_k \,|\, \boldsymbol{x}\right)}{\Pr\left(\mathcal{C}_{|k-1|} \,|\, \boldsymbol{x}\right)} \;\geq\; \frac{\tau'}{1-\tau'} \tag{3.7}$$

$$\Leftrightarrow \quad \Lambda_k(\boldsymbol{x}) = \frac{\mathcal{P}_{\mathcal{C}_k}(\boldsymbol{x})}{\mathcal{P}_{\mathcal{C}_{|k-1|}}(\boldsymbol{x})} \;\geq\; \frac{\tau' \cdot \Pr\left(\mathcal{C}_{|k-1|}\right)}{(1-\tau') \cdot \Pr\left(\mathcal{C}_k\right)} = \tau'', \tag{3.8}$$

or, for reasons of symmetry ($\log z = -\log 1/z$),

$$\Pr\left(\mathcal{C}_k \,|\, \boldsymbol{x}\right) \geq \tau' \quad \Leftrightarrow \quad \log \Lambda_k(\boldsymbol{x}) \;\geq\; \log \tau'' = \tau\,, \tag{3.9}$$

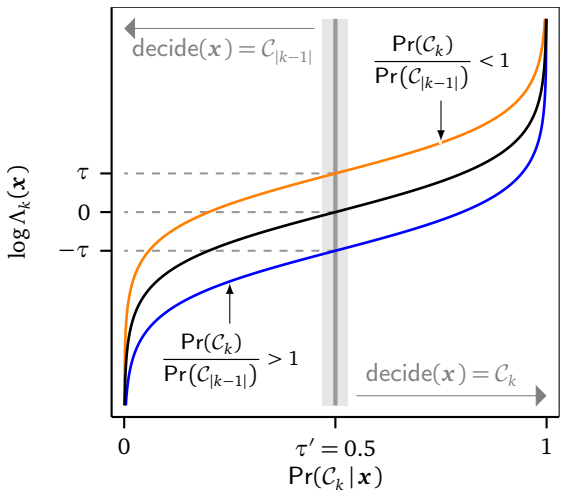


Figure 3.5: A-posteriori probability $\Pr(\mathcal{C}_k \mid \boldsymbol{x})$ and log-likelihood ratio $\log \Lambda_k(\boldsymbol{x})$ in digital image forensics. Each threshold $\Pr(\mathcal{C}_k \mid \boldsymbol{x}) \geq \tau'$ has an equivalent likelihood ratio test, with a threshold parameter τ dependent on the prior class probabilities. The shaded gray area is of width $2\,P_{\text{sep}}$ and alludes to a potential reject option, Equation (3.6).

where Equation (3.8) follows directly from Equation (3.3), and $\Lambda_k(\boldsymbol{x})$ denotes the *likelihood ratio*. Hence, Equations (3.5) to (3.8) suggest that comparing $\Pr(\mathcal{C}_k \mid \boldsymbol{x})$ against a threshold τ' is equivalent to a *likelihood ratio test* with threshold value τ and decision regions

$$\mathscr{R}_k = \{\boldsymbol{x} \mid \log \Lambda_k(\boldsymbol{x}) \geq \tau\} \quad \text{and} \quad \mathscr{R}_{|k-1|} = \{\boldsymbol{x} \mid \log \Lambda_k(\boldsymbol{x}) < \tau\}\,. \tag{3.10}$$

Figure 3.5 illustrates the relation between $\Pr(\mathcal{C}_k \mid \boldsymbol{x})$ and $\log \Lambda_k(\boldsymbol{x})$ and exemplarily marks decision thresholds for $\tau' = 0.5$. The graphs indicate that the evidence for the presence of identifying traces of class $\mathcal{C}_k$ generally decreases as the likelihood ratio becomes smaller. Non-uniform priors shift threshold τ by $\log\left(\Pr(\mathcal{C}_{|k-1|})/\Pr(\mathcal{C}_k)\right)$ and can hence either compensate or amplify low evidence from $\Lambda_k(\boldsymbol{x})$.

The following three examples review fundamental decision rules [115] and give expressions for τ' and τ, respectively.

Example 7 | Minimum Probability of Error. It follows from Bayes' theorem, Equation (3.3), and $\Pr(\mathcal{C}_k) \cdot \mathcal{P}_{\mathcal{C}_k}(\boldsymbol{x}) \propto \Pr(\mathcal{C}_k \mid \boldsymbol{x})$ that the probability of error is minimized by a decision rule that chooses the class with maximum a posteriori probability. The corresponding threshold in Equations (3.5) and (3.7) is $\tau'_{\text{MAP}} = 0.5$, and therefore (see also Figure 3.5)

$$\texttt{decide}_{\text{MAP}}(\boldsymbol{x}) = \mathcal{C}_k \quad \Leftrightarrow \quad \log \Lambda_k(\boldsymbol{x}) \geq \log \frac{\Pr(\mathcal{C}_{|k-1|})}{\Pr(\mathcal{C}_k)} = \tau_{\text{MAP}}\,. \tag{3.11}$$

The subscript 'MAP' hints to the maximum a posteriori rule.

Because it is generally not possible to design algorithms `decide` to reduce both P_M and P_{FA} at the same time, the minimization of Equation (3.4) is often subject to additional constraints. Such constraints may arise from situations where certain misclassifications are considered more critical than others, because of the higher *cost* associated with this type of error (for instance in Example 4, where a false detection of processing artifacts may lead to further time-consuming but unnecessary investigations of the image's authenticity).

Example 8 | Minimum Bayes Risk. Denoting $\gamma^{(l \mapsto k)} \geq 0$ as the cost for assigning an image $\boldsymbol{x}^{(l)}$ to class $\mathcal{C}_k$, it is then the goal to minimize the average cost over all decisions. This is also

known as the *Bayes risk*, $\mathcal{R}$,

$$\mathcal{R} = \sum_{k=0}^{1}\sum_{l=0}^{1} \mathsf{Pr}\left(\mathcal{C}_l\right) \sum_{\boldsymbol{x}\in\mathscr{R}_k} \mathcal{P}_{\mathcal{C}_l}(\boldsymbol{x}) \cdot \gamma^{(l\mapsto k)}\,. \tag{3.12}$$

It follows from Bayes' theorem, that Equation (3.12) is minimized by assigning $\boldsymbol{x}$ to the class $\mathcal{C}_k$ for which the quantity $\sum_l \mathsf{Pr}\left(\mathcal{C}_l \,|\, \boldsymbol{x}\right) \cdot \gamma^{(l\mapsto k)}$ is minimal. Under the reasonable assumption that correct decisions do not cause any cost, $\gamma^{(k\mapsto k)} = 0$, this is equivalent to

$$\begin{aligned}
\texttt{decide}_{\text{MBR}}(\boldsymbol{x}) = \mathcal{C}_k \quad &\Leftrightarrow \quad \mathsf{Pr}\left(\mathcal{C}_k \,|\, \boldsymbol{x}\right) \cdot \gamma^{(k\mapsto|k-1|)} \geq \mathsf{Pr}\left(\mathcal{C}_{|k-1|} \,|\, \boldsymbol{x}\right) \cdot \gamma^{(|k-1|\mapsto k)} && (3.13)\\
&\Leftrightarrow \quad \mathsf{Pr}\left(\mathcal{C}_k \,|\, \boldsymbol{x}\right) \geq \frac{\gamma^{(|k-1|\mapsto k)}}{\gamma^{(|k-1|\mapsto k)} + \gamma^{(k\mapsto|k-1|)}} = \tau'_{\text{MBR}} && (3.14)\\
&\Leftrightarrow \quad \log \Lambda_k(\boldsymbol{x}) \geq \log \frac{\mathsf{Pr}\left(\mathcal{C}_{|k-1|}\right) \cdot \gamma^{(|k-1|\mapsto k)}}{\mathsf{Pr}\left(\mathcal{C}_k\right) \cdot \gamma^{(k\mapsto|k-1|)}} = \tau_{\text{MBR}}\,, && (3.15)
\end{aligned}$$

where the subscript 'MBR' indicates the minimum Bayes risk decision rule. For equal costs $\gamma^{(k\mapsto|k-1|)} = \gamma^{(|k-1|\mapsto k)}$ minimizing $\mathcal{R}$ reduces to minimizing the probability of error.

The above example illustrates that the necessary evidence for the presence of identifying traces of generation functions $(\texttt{acquire}, \texttt{process}) \in (\mathscr{A} \times \mathscr{P})^{(\mathcal{C}_k)}$ increases with the relative cost of misclassifications $\boldsymbol{x}^{(|k-1|)} \mapsto \mathcal{C}_k$. Threshold τ'_{MBR} tends to 1 as false alarms become unaffordable ($\gamma^{(|k-1|\mapsto k)} \to \infty$) and misses are comparably less critical.

On a more general level, the Neyman-Pearson lemma [172] implies for any given probability of false alarm that thresholds values τ and τ' exist, which reduce the probability of miss to a minimum.

Example 9 | Neyman-Pearson. When imposing a bound on the probability of false alarm, $P_{FA} \leq \alpha$, forensic investigators can minimize the probability of miss, $P_{MD}(\alpha)$, by deciding $\mathcal{C}_* = \mathcal{C}_k$ if the likelihood ratio Λ_k is larger than a threshold $\tau''_{\text{NP}}(\alpha)$. This threshold can be found from

$$\alpha = \sum_{\left\{\boldsymbol{x} \,|\, \Lambda_k(\boldsymbol{x}) > \tau''_{\text{NP}}(\alpha)\right\}} \mathcal{P}_{\mathcal{C}_{|k-1|}}(\boldsymbol{x})\,. \tag{3.16}$$

Hence, the Neyman-Pearson (NP) decision rule with $P_{FA} \leq \alpha$ is given by

$$\begin{aligned}
\texttt{decide}_{\text{NP},\alpha}(\boldsymbol{x}) = \mathcal{C}_k \quad &\Leftrightarrow \quad \log \Lambda_k(\boldsymbol{x}) > \log \tau''_{\text{NP}}(\alpha) = \tau_{\text{NP}}(\alpha) && (3.17)\\
&\Leftrightarrow \quad \mathsf{Pr}\left(\mathcal{C}_k \,|\, \boldsymbol{x}\right) > \frac{\tau''_{\text{NP}}(\alpha)}{\frac{\mathsf{Pr}(\mathcal{C}_{|k-1|})}{\mathsf{Pr}(\mathcal{C}_k)} + \tau''_{\text{NP}}(\alpha)} = \tau'_{\text{NP}}(\alpha)\,, && (3.18)
\end{aligned}$$

where Equation (3.18) follows from Equation (3.8). In the classical Neyman-Pearson setting, no assumptions on prior probabilities are made, i. e., $\mathsf{Pr}(\mathcal{C}_0) = \mathsf{Pr}(\mathcal{C}_1)$. Algorithms $\texttt{decide}_{\text{NP},\alpha}$ then minimize the probability of error conditional to the constraint $P_{FA} \leq \alpha$.

Clearly, the decision rules in Examples 8 and 9 can be transformed into each other by setting parameters $\gamma^{(l\mapsto k)}$ and α appropriately. In general, small values α resemble high costs $\gamma^{(|k-1|\mapsto k)}$. Since these parameters are under full control of the forensic investigator, it is furthermore always possible to map problems with non-uniform class priors to equivalent thresholds that are independent of $\mathsf{Pr}(\mathcal{C}_k)$ (this corresponds to the black curve in Figure 3.5).

3.2.2.2 ϵ-Undecidability

Equations (3.7) to (3.9) suggest that the decidability of forensic problems ultimately depends on the mutual similarity of the corresponding class likelihoods $\mathcal{P}_{\mathcal{C}_0}$ and $\mathcal{P}_{\mathcal{C}_1}$, respectively. The more similar these probability distributions are, the more difficult it is to distinguish between samples of them. In the extreme case, when both distributions coincide, i. e., $\log \Lambda(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathbb{X}$, the forensic investigator cannot learn anything from observing $\boldsymbol{x}$.

By inspecting its expected value over all realizations $\boldsymbol{x} \in \mathbb{X}$, the log-likelihood ratio hence reveals insight into the average decidability of digital image forensics problems. This expected value is is also known as *Kullback–Leibler divergence* [136, 42], D_{KL}, and it is given by

$$\mathsf{E}_{\mathcal{P}_{\mathcal{C}_k}}\left(\log \Lambda_k(\boldsymbol{x})\right) = \sum_{\boldsymbol{x}\in\mathbb{X}} \mathcal{P}_{\mathcal{C}_k}(\boldsymbol{x}) \log \frac{\mathcal{P}_{\mathcal{C}_k}(\boldsymbol{x})}{\mathcal{P}_{\mathcal{C}_{|k-1|}}(\boldsymbol{x})} \equiv \mathsf{D}_{\mathrm{KL}}\left(\mathcal{P}_{\mathcal{C}_k}, \mathcal{P}_{\mathcal{C}_{|k-1|}}\right). \tag{3.19}$$

Kullback–Leibler divergence is a fundamental measure of how different two distributions are. It is non-negative, $\mathsf{D}_{\mathrm{KL}}(\mathcal{P}_{\mathcal{C}_k}, \mathcal{P}_{\mathcal{C}_{|k-1|}}) \geq 0$, with equality if and only if the two distributions are identical, and generally asymmetric, $\mathsf{D}_{\mathrm{KL}}(\mathcal{P}_{\mathcal{C}_0}, \mathcal{P}_{\mathcal{C}_1}) \neq \mathsf{D}_{\mathrm{KL}}(\mathcal{P}_{\mathcal{C}_1}, \mathcal{P}_{\mathcal{C}_0})$. A zero KL divergence implies that forensic investigators cannot distinguish between instances of images of either class. In other words, the underlying classification problem becomes only decidable via a decision bias, for instance in terms of non-uniform costs $\gamma^{(l\mapsto k)}$ and/or prior class probabilities.

This leads us to the notion of ϵ-undecidability of a forensic two-class problem and its implications on the error probabilities of algorithms `decide`.

Definition 11 | ϵ-Undecidability. A forensic problem on a binary class space $\mathscr{C} = \{\mathcal{C}_0, \mathcal{C}_1\}$ is *ϵ-undecidable* if $\mathsf{D}_{\mathrm{KL}}(\mathcal{P}_{\mathcal{C}_k}, \mathcal{P}_{\mathcal{C}_{|k-1|}}) \leq \epsilon$ for each $k \in \{0, 1\}$.

Note the similarity of Definition 11 with the notion of ϵ-secure steganography [22]. Also in digital image forensics, ϵ bounds the error probabilities of algorithms `decide` from below via the deterministic processing theorem. More specifically, let $\widetilde{\mathcal{P}}_{\mathcal{C}_k}$ and $\widetilde{\mathcal{P}}_{\mathcal{C}_{|k-1|}}$ denote two binary probability mass functions, which return the probabilities of possible outcomes of `decide` according to the columns of Table 3.1, i. e.,

$$\widetilde{\mathcal{P}}_{\mathcal{C}_k} = \left(\mathsf{Pr}(\texttt{decide} = \mathcal{C}_{|k-1|} \,|\, \mathcal{C}_k), \mathsf{Pr}(\texttt{decide} = \mathcal{C}_k \,|\, \mathcal{C}_k)\right) = \left(P_M, 1 - P_M\right), \tag{3.20}$$

$$\widetilde{\mathcal{P}}_{\mathcal{C}_{|k-1|}} = \left(\mathsf{Pr}(\texttt{decide} = \mathcal{C}_{|k-1|} \,|\, \mathcal{C}_{|k-1|}), \mathsf{Pr}(\texttt{decide} = \mathcal{C}_k \,|\, \mathcal{C}_{|k-1|})\right) = \left(1 - P_{FA}, P_{FA}\right). \tag{3.21}$$

Each of the above distributions corresponds to the (hypothetical) case, where `decide` is fed exclusively instances of images of one particular class $\mathcal{C}_k$ or $\mathcal{C}_{|k-1|}$, respectively. Because the Kullback–Leibler divergence between two probability distributions over $\mathbb{X}$ cannot increase through a measurable map $\boldsymbol{x} \mapsto \texttt{decide}(\boldsymbol{x}) = \mathcal{C}_*$ [136], we have

$$\mathsf{d}_{\mathrm{KL}}\left(P_{FA}, P_{MD}\right) \equiv \mathsf{D}_{\mathrm{KL}}\left(\widetilde{\mathcal{P}}_{\mathcal{C}_{|k-1|}}, \widetilde{\mathcal{P}}_{\mathcal{C}_k}\right) \leq \mathsf{D}_{\mathrm{KL}}\left(\mathcal{P}_{\mathcal{C}_{|k-1|}}, \mathcal{P}_{\mathcal{C}_k}\right), \tag{3.22}$$

where $\mathsf{d}_{\mathrm{KL}}(u, v)$ denotes the Kullback–Leibler divergence between two binary probability distributions $(1-u, u)$ and $(v, 1-v)$, respectively [22].[29] It follows for ϵ-undecidable image

29 Note that, while $\mathsf{D}_{\mathrm{KL}}\left(\mathcal{P}_{\mathcal{C}_{|k-1|}}, \mathcal{P}_{\mathcal{C}_k}\right) = 0$ implies that the underlying forensic problem is undecidable for *all* possible algorithms `decide`, $\mathsf{D}_{\mathrm{KL}}\left(\widetilde{\mathcal{P}}_{\mathcal{C}_{|k-1|}}, \widetilde{\mathcal{P}}_{\mathcal{C}_k}\right) = 0$ only refers to undecidability with one specific decision rule.

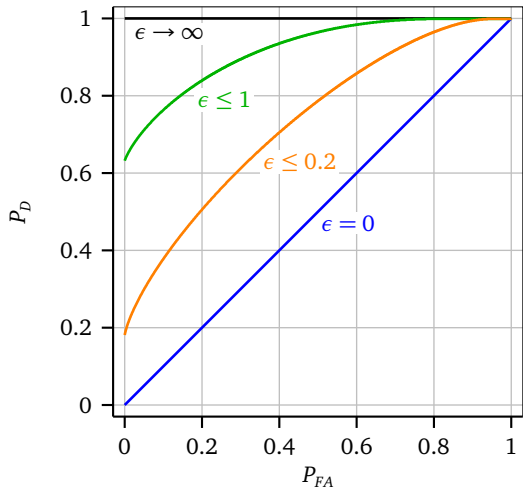


Figure 3.6: Kullback–Leibler divergence upper bounds on the probability of correct detection P_D as a function of the probability of false alarm P_{FA} for algoritms `decide` when examining ϵ-undecidable forensic problems. ROC curves for $\epsilon \in \{0, 0.2, 1, \infty\}$.

forensics problems that the error probabilities of arbitrary digital image forensics algorithms are bounded by

$$0 \leq (1 - P_{FA}) \cdot \log \frac{1 - P_{FA}}{P_M} + P_{FA} \cdot \log \frac{P_{FA}}{1 - P_M} \leq \epsilon \,. \tag{3.23}$$

Equation (3.23) now allows to find the smallest possible probability of missed detection that a forensic algorithm can achieve when examining an ϵ-undecidable problem with a fixed maximum probability of false alarm, $0 \leq P_{FA} \leq 1$,

$$P_M\left(P_{FA}\right) = \arg \min_{P_M \in [0,1]} \left\{ P_M \,\middle|\, \mathrm{d}_{\mathrm{KL}}\left(P_{FA}, P_M\right) \leq \epsilon \right\}, \tag{3.24}$$

where we make use of an equivalent result for steganographic security [66, p. 84].

Figure 3.6 illustrates the resulting upper bounds on the *probability of correct detection*, $P_D = 1 - P_M$, as a function of P_{FA} for various values of ϵ. These *receiver-operating-characteristic* (ROC) curves are particularly indicative for the theoretical limits of Kullback–Leibler divergence. Indistinguishable class likelihoods ($\epsilon = 0$) imply that forensic investigators cannot distinguish between samples of either distribution and hence result in a curve $P_D = P_{FA}$. A maximum divergence ($\epsilon \to \infty$) allows perfect separation, i. e., $P_D = 1$ independent of P_{FA}. In general, it is more likely to correctly identify instances of images of class $\mathcal{C}_k$ at a low probability of false alarms as divergence increases. In very restrictive settings that do not permit any false alarms, the minimum probability of miss decays exponentially with ϵ, $P_M(P_{FA} = 0) \geq \exp(-\epsilon)$.

For decisions according to the maximum a-posteriori rule (Example 7), lower bounds on the probability of error can be found from Toussaint's J-divergence [226], $\mathrm{D}_{\mathrm{J}}^{\pi}$. Denoting $(\pi, 1 - \pi) = (\mathrm{Pr}(\mathcal{C}_0), \mathrm{Pr}(\mathcal{C}_1))$, this measure generalizes Kullback and Leibler's [136] symmetric J-divergence,

$$\mathrm{D}_{\mathrm{J}}^{\pi}\left(\mathcal{P}_{\mathcal{C}_0}, \mathcal{P}_{\mathcal{C}_1}\right) = \sum_{\boldsymbol{x} \in \mathbb{X}} \left(\pi \cdot \mathcal{P}_{\mathcal{C}_0}(\boldsymbol{x}) - (1 - \pi) \cdot \mathcal{P}_{\mathcal{C}_1}(\boldsymbol{x})\right) \log \frac{\pi \cdot \mathcal{P}_{\mathcal{C}_0}(\boldsymbol{x})}{(1 - \pi) \cdot \mathcal{P}_{\mathcal{C}_1}(\boldsymbol{x})} \tag{3.25}$$

$$= \pi \cdot \mathrm{D}_{\mathrm{KL}}\left(\mathcal{P}_{\mathcal{C}_0}, \mathcal{P}_{\mathcal{C}_1}\right) + (1 - \pi) \cdot \mathrm{D}_{\mathrm{KL}}\left(\mathcal{P}_{\mathcal{C}_1}, \mathcal{P}_{\mathcal{C}_0}\right) + \mathrm{d}_{\mathrm{KL}}(\pi, \pi) \,. \tag{3.26}$$

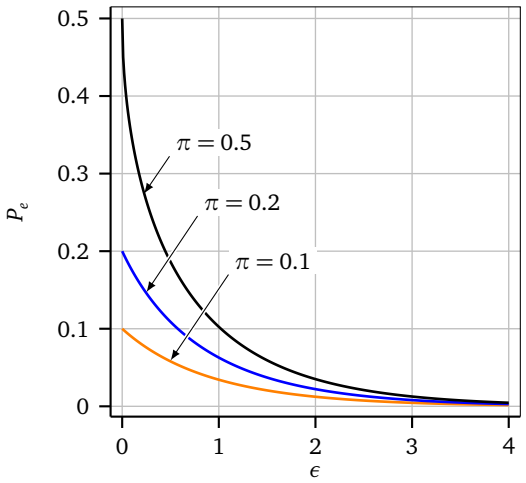


Figure 3.7: J-divergence lower bounds on the probability of error P_e for algorithms $\texttt{decide}_{\text{MAP}}$ when examining ϵ-undecidable problems. Curves for various class prior probabilities $\pi = \Pr(\mathcal{C}_0)$, $\pi \in \{0.5, 0.2, 0.1\}$.

$\mathsf{D}_{\mathrm{J}}^{\pi}$ is shown in [226] to bound the probability of error of MAP decision rules by

$$P_e \geq 0.5 - 0.5 \cdot \sqrt{1 - 4 \exp\left(-2\,\mathsf{H}(\pi, 1-\pi) - \mathsf{D}_{\mathrm{J}}^{\pi}\left(\mathcal{P}_{\mathcal{C}_0}, \mathcal{P}_{\mathcal{C}_1}\right)\right)}, \tag{3.27}$$

where H denotes the Shannon entropy. It further follows from Equation (3.26) and Definition 11 that ϵ-undecidable problems are characterized by

$$\mathsf{D}_{\mathrm{J}}^{\pi}\left(\mathcal{P}_{\mathcal{C}_0}, \mathcal{P}_{\mathcal{C}_1}\right) \leq \epsilon + \mathsf{d}_{\mathrm{KL}}(\pi, \pi)\,. \tag{3.28}$$

Combining Equations (3.27) and (3.28), we obtain the following bound on the performance of algorithms $\texttt{decide}_{\text{MAP}}$:

$$P_e \geq 0.5 - 0.5 \cdot \sqrt{1 - 4\pi(1-\pi) \cdot \exp(-\epsilon)}\,. \tag{3.29}$$

Figure 3.7 illustrates this bound for different class prior probabilities and indicates that priors with $\pi \neq 0.5$ reduce the probability of error. The curves further suggest that non-uniform priors are particularly influential when the corresponding class likelihoods are highly similar ($\epsilon \to 0$), which conforms to our earlier discussion on page 48.

We finally note that Toussaint's bound can be generalized to $K > 2$ classes [227]. Moreover, also upper bounds on P_e exist, most prominently the Chernoff bound [98] and the Bhattacharyya bound [113]. However, as shown by van Ness [171], non-trivial upper bounds in terms of $\mathsf{D}_{\mathrm{KL}}\left(\mathcal{P}_{\mathcal{C}_k}, \mathcal{P}_{\mathcal{C}_{|k-1|}}\right)$ cannot exist.

3.2.2.3 Multiple Images

So far, we have considered the forensic examination of single images. In certain situations, forensic investigators may be in the position to analyze a set of (presumably) independent images $\mathbb{X}_n = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{n-1}\}$, which (are assumed to) result from the same set of generation functions $(\mathscr{A} \times \mathscr{P})^{(C)}$. Again, it is then the goal to assign $\mathbb{X}_n$ to a class $\mathcal{C}_*$ according to some optimality criterion, and the above decision rules basically apply in a straightforward manner

by evaluating $\mathcal{P}_{\mathcal{C}}(\mathbb{X}_n)$ instead. A typical application is source device identification of a batch of related images [3].[30]

Intuitively, we expect that forensic investigators will make more reliable decisions as the number of available images, n, increases. For the Neyman-Pearson decision rule and independent samples $\boldsymbol{x}_i \sim \mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$, this behavior is captured by the Chernoff-Stein Lemma [42, ch. 11.8]. The lemma states that, for any fixed probability of false alarm, $P_{FA} \leq \alpha$, the best achievable probability of missed detection, $P_M(\alpha)$, decays exponentially with n, i. e.,

$$\lim_{n\to\infty} \frac{1}{n} \log P_M(\alpha) = -\,\mathsf{D}_{\mathrm{KL}}\left(\mathcal{P}_{\mathcal{C}_{|k-1|}}, \mathcal{P}_{\mathcal{C}_k}\right). \tag{3.30}$$

Observe that, via Kullback–Leibler divergence, the above equation establishes a direct link to the notion of ϵ-reliability. A similar result exists for MAP decision rules [42, ch. 11.9],

$$\lim_{n\to\infty} \frac{1}{n} \log P_e = -\mathsf{C}\left(\mathcal{P}_{\mathcal{C}_k}, \mathcal{P}_{\mathcal{C}_{|k-1|}}\right), \tag{3.31}$$

where $\mathsf{C}(\mathcal{P}_{\mathcal{C}_k}, \mathcal{P}_{\mathcal{C}_{|k-1|}})$ is the Chernoff information,

$$\mathsf{C}\left(\mathcal{P}_k, \mathcal{P}_{|k-1|}\right) = \inf_{0<\lambda<1} \log \sum_{\boldsymbol{x}\in\mathbb{X}} \mathcal{P}^{\lambda}_{\mathcal{C}_k}(\boldsymbol{x}) \cdot \mathcal{P}^{1-\lambda}_{\mathcal{C}_{|k-1|}}(\boldsymbol{x}). \tag{3.32}$$

As pointed out by Cover and Thomas [42, pp. 388–389], the asymptotic bound in Equation (3.31) does not depend on the class prior probabilities, i. e., the effect of prior knowledge washes out for large sample sizes for strictly positive Chernoff information.

3.3 Practical Considerations

The theory in Sections 3.1 and 3.2 provides a framework general enough to cover and discuss a wide range of questions regarding digital image forensics. Unfortunately, only a few of the definitions are directly applicable in practice. Knowledge of the class likelihoods $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$ is the key to optimal decisions conditional on the class prior probabilities $\mathsf{Pr}(\mathcal{C})$. Both quantities are however hardly ever available in practical settings. Insufficient knowledge of prior probabilities is less problematic:[31] the Neyman-Pearson lemma allows to make optimal decisions with respect to a chosen probability of false alarm independent of prior probabilities, see Equation (3.17). The main difficulty in applying the above equations remains the absence of knowledge about the conditional probability distributions $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$, which are given only empirically and are of intractable dimensionality (Section 3.3.1). Forensic investigators hence need to find models of digital images (Section 3.3.2). Moreover, the question whether or not an image is authentic is hard to evaluate in practice because of the empirical nature of semantic equivalence and the high complexity of image generation functions `generate` (Section 3.3.3).

30 Also certain forensic analyses of digital video footage can be interpreted as examination of multiple images of the same source. However, due to temporal correlation, individual frames are typically not independent, which can complicate a formal treatment. On the other hand, it has already been demonstrated that these dependencies provide useful information for forensic purposes [232, 233, 101, i. a.]

31 Reasonable (or conservative) assumptions on $\mathsf{Pr}(\mathcal{C})$ may exist when the class space reflects different (types of) imaging devices.

3.3.1 Epistemic Bounds

We follow Böhme [18, pp. 83–86], who transferred widely accepted epistemological paradigms to the context of steganography and steganalysis and argued that distributions $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$ are ultimately incognizable. Although the support of $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$ is finite, these class likelihoods return probability values of projections of real-world phenomena $\mathcal{S} \in \mathscr{S}$, and $\mathscr{S}$ has infinite support (so has the parameter space Θ), cf. Definitions 1 and 9. Because natural phenomena in the real world can never be fully known but merely approximated by consequent falsification and refinement of theories about the real world, complete knowledge of $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$ remains inaccessible to the forensic investigator.

But even after the transformation to the finite space $\mathbb{X}$, the support of $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$ is too large and too heterogeneous to efficiently estimate distributions by sampling. We can ignore this for a moment and assume that original (or authentic) images can efficiently be sampled.[32] But then there remains the difficulty of sampling inauthentic images. Generating good counterfeits is a time-consuming *manual* task that depends a lot on the counterfeiters' creativity and it is highly adaptive to the original image. This process is generally very hard to automate, although promising semi-automatic approaches have been proposed recently [140, 50].

The high cost of sampling is also reflected by the quality and size of available datasets that have been compiled in controlled environments for the purpose of image forensics research. Typical inauthentic images are obtained by copying patches from within the same or other images, without sophisticated post-processing and without adaptivity to the depicted scene [175, 102]. Only few databases provide more realistic forgeries [36, 39], however without resolving the general trade-off between quality and quantity.[33]

3.3.2 Image Models

To reduce complexity and avoid the epistemic obstacles, all practical digital image forensics algorithms make use of *models* of digital images. Such models can be seen as a dimensionality reduction by projecting the high-dimensional image space $\mathbb{X}$ to a much smaller and more tractable subspace, which is usually referred to as *feature space*.

Definition 12 | Feature Representation. Digital images $\boldsymbol{x} \in \mathbb{X} \equiv \mathscr{X}^N$ are mapped to feature vectors $\boldsymbol{y} \in \mathbb{Y} \equiv \mathscr{Y}^{N'}$ by a function $\texttt{project} : \mathbb{X} \to \mathbb{Y}$. For a given class $\mathcal{C} \in \mathscr{C}$, the class likelihood over the feature space $\mathbb{Y}$, $\mathcal{P}_{\mathcal{C}}^{\texttt{project}}(\boldsymbol{y})$, is given via the class likelihood $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$:

$$\mathcal{P}_{\mathcal{C}}^{\texttt{project}}(\boldsymbol{y}) = \sum_{\{\boldsymbol{x} \,|\, \boldsymbol{x} = \texttt{project}^{-1}(\boldsymbol{y})\}} \mathcal{P}_{\mathcal{C}}(\boldsymbol{x})\,. \tag{3.33}$$

32 In practice, it should be distinguished between large-scale sampling from (a moderate number of) particular devices or all possible (types of) devices. The former is to a certain degree feasible [82], but it is unrealistic to assume that forensic investigators are granted access to arbitrary instances of `acquire`. This is of particular relevance to the identification of specific source devices, where images of unknown devices may occur with non-negligible probability.

33 Ng et al. [181, p. 402] suspected that the main reason for the lack of realistic inauthentic test images lies in the inferior sophistication of forensic algorithms. We respond that imperfect algorithms should not hinder forensic investigators to compile more realistic datasets. It is ultimately the only way to gain knowledge about distributions $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$, which in turn will eventually advance forensic algorithms.

Convention. To simplify notation, we use $\mathcal{P}_{\mathcal{C}}(\boldsymbol{y})$ as a shortcut to $\mathcal{P}_{\mathcal{C}}^{\texttt{project}}(\boldsymbol{y})$, whenever the context prevents ambiguities.

In the simplest case, function `project` maps images $\boldsymbol{x}$ to scalar values y, i. e., $\mathbb{Y} = \mathbb{R}$. Typical examples are the energy of high frequency components of pixel histograms in the detection of contrast enhancement [211], or the ratio of noise energies in horizontal and vertical direction in the distinction between images from scanners and cameras [23]. We note that a projection to the feature space $\mathbb{Y}$ not necessarily means that $N' \ll N$. For instance, most PRNU-based methods project images to noise vectors $\boldsymbol{y} = (y_1, \ldots, y_N)$, which are of the same length as the images itself [65].

The projection of images to a feature space does generally not relieve forensic investigators of their epistemic bounds: distributions $\mathcal{P}_{\mathcal{C}}(\boldsymbol{y})$, via the class likelihoods $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$, still depend on the incognizable distribution $\mathcal{S} \sim \mathscr{S}$, cf. Equation (3.33). However, an appropriate choice of functions `project` allows to formulate reasonable and mathematically tractable assumptions on the distribution of images in the feature space. A common assumption for noise vectors, for instance, is the statistical independence of their elements. Due to the high dimensionality and complexity of digital images, such assumptions are typically not viable for distributions $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$ directly. Böhme [18, p. 85] calls such *image models* hypotheses on $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$, because assumptions on the distribution of feature vectors, $\mathcal{P}_{\mathcal{C}}(\boldsymbol{y})$, indirectly translate to assumptions on the class likelihoods over $\mathbb{X}$.

Example 10 | PRNU-based Device Identification (see Section 2.6.1.2). Images are mapped to the feature space by extracting a noise residual, $\boldsymbol{y} = \boldsymbol{x} - \mathsf{denoise}(\boldsymbol{x})$, where $\mathsf{denoise} : \mathbb{X} \to \mathbb{R}^N$ is a denoising filter. Classes $\mathcal{C}_k$ are defined to entail instances of images of particular devices k, cf. Example 3. In testing whether an image was captured with device k, individual noise samples y_i are modeled by a Gaussian distribution, and it is assumed that a common PRNU term $\boldsymbol{x}\boldsymbol{\kappa}^{(k)}$ controls the distribution $\mathcal{P}_{\mathcal{C}_k}$, i. e., $\mathcal{P}_{\mathcal{C}_k}(y_i) = \mathcal{N}\big(x_i\kappa_i^{(k)}, \sigma_i^2\big)$. Images of devices $l \neq k$ do not exhibit this identifying characteristic, and hence $\mathcal{P}_{\mathcal{C}_l}(y_i) = \mathcal{N}(0, \sigma_i^2)$.

Example 11 | Copy–Move Detection (see Section 2.6.2.1). Small blocks of the image are mapped to a feature representation that is robust against noise and other distortions. All transformed blocks are then compared to each other, leading to a low-resolution detection map $\boldsymbol{y}$, which marks pairs of (near-)duplicate blocks. The implicit model assumes that connected regions of identical, but not constant, pixel values are very unlikely to occur in original images.

The two examples highlight the great variety in ways of specifying image models. While the former builds on rigorous mathematical modeling, the latter is implicit and of rather informal kind. In general, modeling images in low-dimensional feature spaces is effective as long as the mismatch with the real world is not substantial. By accepting the general need for image models, it is clear that a forensic algorithm can only be as good as the model it employs. The better the underlying model can explain and predict observed samples of a particular class, the more confident a forensic investigator can base her decisions on it.

Depending on the available knowledge of the distribution of samples $\boldsymbol{y}$ in the feature space, different options exist for incorporating these models into the decision stage. More generally, we can identify the following approaches to formulate decision rules in terms of image models [14, p. 43]:

|1| *Generative models* are explicit models of distributions $\mathcal{P}_{\mathcal{C}}(\boldsymbol{y})$.[34] If available for all classes $\mathcal{C}_k \in \mathscr{C}$, this type of model is generally preferable, as it establishes a direct link to the class likelihoods $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$. In the theory of Section 3.2.2, intractable distributions over $\mathbb{X}$ are then replaced by distributions over the codomain $\mathbb{Y}$, which allows to find optimal decision rules with respect to the model.

|2| Instead of modeling $\mathcal{P}_{\mathcal{C}}(\boldsymbol{y})$ directly, *discriminative models* make implicit assumptions on the class likelihoods via models of the posterior $\Pr(\mathcal{C} \mid \boldsymbol{y})$. It is then possible to solve the decision problem in Equation (3.5) by imposing appropriate thresholds on $\Pr(\mathcal{C} \mid \boldsymbol{y})$ directly. While this approach is in general not as universal as the explicit modeling of $\mathcal{P}_{\mathcal{C}}(\boldsymbol{y})$ (it is always possible to obtain posterior probabilities from the class likelihoods via Bayes' theorem), it has particular merits when reasonable assumptions on (some of the) class likelihoods are not available.

|3| An even simpler approach employs *discriminant functions*, which directly map each sample $\boldsymbol{y}$ to a class $\mathcal{C}_k \in \mathscr{C}$. Here, no probabilities are involved, and models of the distribution of feature values are again implicit only. This approach is particularly attractive when forensic investigators are only interested in 'hard' decisions.

We note that even after projection to the feature space, explicit models of class likelihoods often still do not exist. As a result, only few forensic methods make use of generative models (most prominently, device identification via sensor noise, cf. Example 10). The vast majority of algorithms employs discriminant functions in the feature space to assign images to particular classes. Projections to one-dimensional feature values are usually accompanied by empirical decision thresholds [23, 211, i. a.], whereas feature representations of higher dimensionality are mostly fed into support vector machines [41] to find the maximum margin decision boundary [28, 85, i. a.]. To the best of our knowledge, discriminative models have not found direct application in the context of digital image forensics yet. Nevertheless, it seems straightforward to extend simple discriminant functions to return soft probability values, which may support forensic investigators in making informed decisions. Swaminathan et al.'s [220] use of probabilistic support vector machines [192, 238] to obtain 'confidence scores' is a step in this direction.[35] More sophisticated methods, which do not rely on a proxy discriminant function, include for instance the relevance vector machine [225].

Independent of how complex image models are and how they are eventually incorporated into the decision stage, all good models have in common that they are ideally inferred from and/or validated against a large number of representative samples. If image models are inferred from observed samples, *training sets* of limited size and lacking heterogeneity are likely to result in models that do not generalize well to images from other sources. Similarly, explicit (possibly analytical) assumptions on class-specific characteristics of feature vectors need to be tested against a representative *validation set* to rule out a mismatch with the real world. These points

34 The term 'generative' reflects the possibility to sample from $\mathcal{P}_{\mathcal{C}}(\boldsymbol{y})$ [14].

35 It is worth mentioning that a considerable portion of practical image forensics methods does not output decision scores at all. This mainly concerns methods designed to uncover local manipulations by analyzing images on a block-by-block basis (for instance copy-move detectors, cf. Example 11). The standard output is a detection map $\boldsymbol{y}$ that points to conspicuous image regions, and decisions typically require human interpretation (see Figure 2.14 for a specific example). It remains unclear whether this raw presentation of detection results is driven by the lack of reasonable models of $\boldsymbol{y}$, or by overly simplistic assumptions thereon. While one could argue that the existence of a single marked block is sufficient to flag the whole image as inauthentic, this approach would ignore important (yet more complex) available information, such as size, shape and connectedness of the conspicuous regions.

emphasizes the major issues that arise from the high cost of sampling inauthentic images, cf. Section 3.3.1.

3.3.3 Blurred Notion of Authenticity

3.3.3.1 Measurement of Semantic Equivalence

Epistemic bounds not only limit the forensic investigator's knowledge of class likelihoods, but they further directly affect the assessment of image authenticity. While image models are generally a viable option to resolve the former issue, the direct dependence on $\mathcal{S}$ prevents a literal application of Equation (3.1) to test for the semantic equivalence of digital images $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ in Definition 7. A possible strategy to work around this difficulty is to interpret function $\mathsf{semantic.dist}$ as a conditional probability distribution $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}_1 \,|\, \mathcal{S})$, and to assume that one of the images is a satisfactory representation of natural phenomenon $\mathcal{S}$. It is then possible to link function $\mathsf{semantic.dist}$ to assumptions on the distribution of images by finding a model for distribution $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}_2 \,|\, \boldsymbol{x}_1)$. This procedure is apparently only feasible if the model is sensitive enough to capture semantic differences between the images, and it remains an open question whether such models exist. A very coarse approximation of $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}_2 \,|\, \boldsymbol{x}_1)$ may be obtained by using a (much simpler) function $\mathsf{dist} : \mathbb{X} \times \mathbb{X} \to \mathbb{R}_+$ to compute differences between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ on sample-level, and to assume that large differences reflect low probabilities,

$$\Big|\mathsf{semantic.dist}(\boldsymbol{x}_1, \mathcal{S}) - \mathsf{semantic.dist}(\boldsymbol{x}_2, \mathcal{S})\Big| \propto \mathsf{dist}(\boldsymbol{x}_1, \boldsymbol{x}_2) \tag{3.34}$$

Among the simplest distance functions of this kind is the peak-signal-to-noise ratio (PSNR),

$$\mathsf{dist}_{\mathrm{PSNR}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = 10 \cdot \log_{10} \left(\frac{|\boldsymbol{x}_1| \cdot \left(2^{\ell} - 1\right)^2}{\sum_i \left((x_1)_i - (x_2)_i\right)^2} \right) , \tag{3.35}$$

which is measured in decibel (dB). We note that critics have expressed concern that PSNR and related pixel-wise distortion metrics [51] do not reflect very well the way that human visual systems (HVS) perceive images [80, 234]. As a result, a plethora of HVS-inspired alternative distortion measures have been proposed, with the structural similarity (SSIM) index being one of the most promising approaches [235].

3.3.3.2 Legitimate Post-Processing

A further obstacle to the assessment of image authenticity lies in the absence of knowledge of the original image $\boldsymbol{x}^{(1)}$. Without access to $\mathcal{S}$ *and* $\boldsymbol{x}^{(1)}$, forensic investigators can ultimately never know whether arbitrary images $\boldsymbol{x}$ are authentic or not. Consequently, they can only resort to inference on the parameters of function `generate` and apply heuristics that determine which types of processing preserve the semantic meaning of digital images. In practical settings it is hence more appropriate to speak of *plausible images* instead.

While it is tempting to deny authenticity to every processed image *per se*, this simplification is too narrow for many realistic applications. More likely, there will be a subset $\mathscr{P}_{\mathrm{legitimate}} \subset \mathscr{P}$ of *legitimate processing* operations which do not impair the authenticity of an image, see also Figure 3.4. This subset (and thus the notion of authenticity/plausibility) certainly depends on the context. For instance, it is common practice to downscale and compress digital images with

JPEG prior to publication on the Internet [216]. Official crime scene photographs, by contrast, must not be exposed to any quality reduction. Instead, basic color enhancement operations may be accepted or even required [202]. Special codes of conduct exist that specify what is considered legitimate post-processing for a number of scientific journals [205, 186, 44].

Practical investigations into the authenticity of digital images will hence ultimately deal with three categories of digital images, namely

|1| *original images*, where `process` can be nothing but the identity function $\perp_{\mathscr{P}}$,
|2| *plausible images*, which have been subject to legitimate post-processing $\texttt{process} \in \mathscr{P}_{\text{legitimate}}$, and
|3| *manipulated images*, for processing with all other elements of $\mathscr{P}$.

It is then the forensic investigator's goal to distinguish between instances of images of the three categories, and in particular between those that underwent legitimate and illegitimate processing (categories 2 and 3). Context and established habits define whether the first two categories or just the first category shall be considered as authentic.

Example 12 | Context-Dependent Legitimacy. Imagine a case where a judge who has to rule on a traffic accident may consider JPEG-compressed images as authentic if they have been mailed to the insurance company via email. Since the authenticity (and in particular the semantic integrity) of JPEG-compressed images is more difficult to prove than of never-compressed images (cf. Figure 3.3), a party in doubt may present (or demand) the original raw files. The claim "these are the original raw files" immediately alters the notion of authenticity. JPEG artifacts in the presumably never-compressed files would be an indication of inauthenticity and raise suspicion that the images are counterfeits.

Remark that technically, this claim imposes an exogenous condition on the class likelihood $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}\,|\,\text{claim})$. This way, contextual knowledge can be incorporated in the formal framework and sharpen the notion of plausibility with probability distributions.

3.3.3.3 Device-Internal Processing

The definition of authenticity is deeply entangled with the question of what constitutes an acquisition device and thus the separation between functions `acquire` and `process`. Common sense suggests to equate function `acquire` with imaging devices and then strictly apply Definition 7: all original images captured with these devices are authentic. However, considering the sophistication and complexity of modern imaging devices, the situation is not as easy.

Increasingly, such post-processing and image enhancement becomes integral part to the *internal* imaging pipeline and blurs the distinction between functions `acquire` and `process`. The situation is even more complex when consumers are in the position to actively modify and extend the firmware of their devices.[36] Eventually, forensic investigators have to accept that device-internal processing will often be hardly distinguishable from post-processing outside the device and hence raises the uncertainty of forensic decisions.

36 The Canon Hack Development Kit, for instance, allows to run virtually arbitrary image processing routines inside most modern Canon digital cameras, `http://chdk.wikia.com/wiki/CHDK`.

3.4 Counter-Forensics

A thorough assessment of the reliability of digital image forensics algorithms requires to anticipate strategies of potential counterfeiters. Whenever an image is manipulated purposely (and with the intention to make it public to a certain group of entities), the counterfeiter will have at least a rough working definition of what is considered plausible and will try to conform to these expectations. Because knowledge of identifying traces of image generation functions is in general not limited to the forensic investigator, informed counterfeiters will eventually exploit their own knowledge (or, to be more precise: assumptions) to deceive forensic analyses, i. e., to influence the outcomes of digital image forensics algorithms (Section 3.4.1). For a characterization of such *attacks* it is important to distinguish between the *robustness* and the *security* of digital image forensics algorithms, respectively (Section 3.4.2). Because we expect counterfeiters to possess different skills and to have access to different resources, it seems feasible to adopt the notion of *adversary models* to study the reliability of image forensics algorithms conditional to those expectations (Section 3.4.3) and to the potential strategies that counterfeiters can pursue (Section 3.4.4).

3.4.1 Formal Definition

For a given image $\boldsymbol{x}_1 = \boldsymbol{x}_1^{(k)}$, counter-forensics aims at preventing the assignment to the image's class $\mathcal{C}_k$. By suppressing or synthesizing identifying traces, the counterfeiter creates a *counterfeit* $\boldsymbol{x}_2 = \boldsymbol{x}_2^{(\tilde{l})}$ with the intention to let it appear like a plausible member of an alternative class $\mathcal{C}_l \in \mathscr{C}$, $l \neq k$, when presented to the forensic invesitgator's function `decide`.

Convention. We use the superscript notation $^{(\tilde{l})}$ to denote the intended class change.

Definition 13 | Counter-Forensic Attack. A digital image forensics algorithm `decide` is *vulnerable* to a *counter-forensic attack* if for a given image $\boldsymbol{x}_1 = \texttt{generate}(\boldsymbol{\theta})$

$$\exists\, \texttt{attack} \in \mathscr{P},\ \boldsymbol{x}_2 = \texttt{attack}(\boldsymbol{x}_1) \quad \text{so that} \quad \texttt{decide}\left(\boldsymbol{x}_2\right) \neq \texttt{decide}(\boldsymbol{x}_1) \tag{3.36}$$

subject to the constraints

|1| $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are semantically equivalent (*semantic constraint*), and

|2| the probability of finding `attack` for a given $\boldsymbol{x}_1$ is not negligible within a given complexity bound (*computational constraint*).

The following examples illustrate how this definition matches existing counter-forensic strategies (cf. Table 2.2).

Example 13 | Hiding Traces of Image Manipulation. A typical counter-forensic image manipulation of an authentic image $\boldsymbol{x}_1^{(1)}$ will involve two steps, first a transformation $\boldsymbol{x}_1^{(1)} \mapsto \boldsymbol{x}_2^{(0)}$ which changes the semantic according to the counterfeiter's intention, and second a counter-forensic attack $\boldsymbol{x}_2^{(0)} \mapsto \boldsymbol{x}_3^{(\tilde{1})}$ to pretend authenticity of the counterfeit, i. e., $\texttt{decide}(\boldsymbol{x}_3^{(\tilde{1})}) = \mathcal{C}_1$. Images $\boldsymbol{x}_2^{(0)}$ and $\boldsymbol{x}_3^{(\tilde{1})}$ are semantically equivalent.

Example 14 | Impeding Device Identification. Counterfeiting the source of an authentic image $\boldsymbol{x}_1$ involves a single application of a counter-forensic attack $\boldsymbol{x}_1 \mapsto \boldsymbol{x}_2$, possibly with the additional requirement that a specific target class $\mathcal{C}_{\text{target}} \overset{!}{=} \texttt{decide}(\boldsymbol{x}_2) \neq \texttt{decide}(\boldsymbol{x}_1)$ is pretended.

While Definition 13 is formulated for a specific choice of function `decide`, it is also meaningful to study the *reliability* of a counter-forensic attack, which addresses identifying traces of original class $\mathcal{C}$, against all possible decision rules on a given class space $\mathscr{C}$. From the counterfeiter's point view, every forensic analysis can be reduced to a two-class decision problem on a class space $\mathscr{C}' = \{\mathcal{C}'_0, \mathcal{C}'_1\}$ by defining classes $\mathcal{C}'_1$ and $\mathcal{C}'_0 \equiv \mathcal{C}_{\bar{1}}$ to represent the set of target (i. e., admissible) generation functions and attacks, respectively:

$$(\mathscr{A} \times \mathscr{P})^{(\mathcal{C}'_0)} = (\mathscr{A} \times \mathscr{P})^{(\mathcal{C})} \times \{\texttt{attack}\} \tag{3.37}$$

$$(\mathscr{A} \times \mathscr{P})^{(\mathcal{C}'_1)} \subseteq (\mathscr{A} \times \mathscr{P})^{(\mathcal{Q})}\,. \tag{3.38}$$

The concrete definition of class $\mathcal{C}'_1$ depends on the counterfeiter's agenda. It may correspond to a combination of several classes (if the goal is only to suppress identifying traces of class $\mathcal{C}$) or to a particular class $\mathcal{C}_{\text{target}}$ (for instance to pretend a specific source device, cf. Example 14). Careful counterfeiters in general strive to design their attacks so that samples of both classes $\mathcal{C}'_0$ and $\mathcal{C}'_1$ are indistinguishable, and counter-forensic attacks are clearly the more reliable the more similar the corresponding class likelihoods $\mathcal{P}_{\mathcal{C}'_0}(\boldsymbol{x})$ and $\mathcal{P}_{\mathcal{C}'_1}(\boldsymbol{x})$ are. Hence, the reliability of a counter-forensic attack directly relates to the decidability of the forensic problem to distinguish between the two distributions.

Definition 14 | ϵ-Reliability. A counter-forensic attack, which addresses identifying traces of original class $\mathcal{C} \in \mathscr{C}$, is *$\epsilon$-reliable* against all digital image forensics algorithms on $\mathscr{C}$ if the forensic investigator's decision problem to distinguish between samples $\boldsymbol{x}^{(\mathcal{C}'_0)} \sim \mathcal{P}_{\mathcal{C}'_0}$ and $\boldsymbol{x}^{(\mathcal{C}'_1)} \sim \mathcal{P}_{\mathcal{C}'_1}$ is ϵ-undecidable.

For the special case $\epsilon = 0$, Definition 11 implies that authentic and counterfeit images are drawn from the same distribution and the forensic investigator cannot gain any information from the analysis of $\boldsymbol{x}$. Inspired by the related notion of perfect security of steganographic schemes [22, 66], we call such counter-forensic attacks *perfectly reliable*.

3.4.2 Robustness and Security of Image Forensics Algorithms

The design of counter-forensic techniques is not only of practical interest to the counterfeiter itself. It rather also has a strong academic perspective, as it allows to study the security of forensic algorithms `decide`. Before we define what exactly we mean by security, it is useful to come back to the distinction between legitimate and illegitimate post-processing (see Section 3.3.3.2) and thereby to introduce the notion of robustness first.

3.4.2.1 Robustness

Definition 15 | Robustness. The *robustness* of a digital image forensics algorithms is defined by its reliability under legitimate post-processing.

Forensic investigators generally wish to operate highly robust forensic algorithms, which are barely sensitive to any form of legitimate post-processing. This is so because the lack of a clear separation between authentic and inauthentic images increases the counterfeiter's set of strategic options. If quality reduction, such as lossy compression or downscaling, is considered plausible and thus inconspicuous, a counterfeiter can always try to eliminate subtle identifying traces of the original class $\mathcal{C}$ by reducing the semantic resolution of images $\boldsymbol{x}^{(\mathcal{C})}$. The practical

relevance of this approach can be readily seen from the visual quality of typical image forgeries published on the Internet. Indeed, many known forensic algorithms are sensitive to strong quantization. Yet some exceptions exist. For example, scans of printed and dithered images in newspapers are coarse digital representations of the real world, but traces of inconsistent lighting may still be detectable [112].

As a counter-forensic technique, legitimate post-processing does not require much knowledge of the image generation process. Its sole objective is to generate plausible counterfeits. It is sufficient if the counterfeit is moved *somewhere* outside the decision region $\mathscr{R}_{\mathcal{C}}$ (subject to the constraints in Definition 13).

As a consequence, the experimental literature is mainly concerned about the robustness of novel forensic algorithms. Most authors measure and report the performance loss as a function of JPEG compression quality or strength of additive white noise. While this is a good indicator of the *average* reliability, it does not permit conclusions on the overall reliability. A more complete view also has to consider *worst-case* scenarios with sophisticated and intentional counterfeiters. Resistance against such attacks is directly associated with the security of forensic algorithms.

3.4.2.2 Security

Definition 16 | Security. The *security* of a digital image forensics algorithm is defined by its reliability to detect intentionally concealed class changes. In other words, security is the ability to withstand counter-forensics.

Because complete knowledge of the theoretical class likelihoods $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$ is never reachable, the security of forensic algorithms directly relates to the forensic investigator's model of digital images (or, equivalently, the assumptions about $\mathcal{P}_{\mathcal{C}}(\boldsymbol{y})$). By exploiting shortcomings of this model, counterfeiters purposely move the counterfeits *in a particular direction* towards the decision boundary of the original class $\mathcal{C}$ (and just beyond). The more restricted this model is, the easier a counterfeiter can in general find ways to construct successful counter-forensic techniques. In the light of this important observation, counter-forensic techniques clearly benefit from the modus operandi of using low-dimensional projections when assigning digital images to particular classes. A viable and straightforward strategy to increase model dimensionality is to combine several low-dimensional algorithms, which ideally rely on independent feature spaces. This 'suite of detection tools' is exactly what has been propagated by Popescu and Farid [196] (cf. Section **??**). As the dimensionality (i. e, the number of employed forensic algorithms) grows, it becomes increasingly harder for the counterfeiter to move an image into the target decision region in each dimension at the same time.

Moreover, counterfeiters are generally subject to similar practical limitations as forensic investigators. In particular, the very same epistemic bounds apply to the design of both image forensics algorithms `decide` and counter-forensic methods `attack`, respectively. Both sides need to resort to relatively low-dimensional image models, which are ultimately only assumptions about reality. This implies that counterfeiters can never gain perfect knowledge whether their image model is good enough so that no decision function can discriminate between authentic and counterfeit images. The theoretical reliability in Definition 14 cannot be calculated in the absence of knowledge of $\mathcal{P}_{\mathcal{C}'_0}(\boldsymbol{x})$ and $\mathcal{P}_{\mathcal{C}'_1}(\boldsymbol{x})$. A counter-forensic attack will only be successful as long as image forensics algorithms have not been refined accordingly.

The best that counterfeiters can hope for is perfect reliability with respect to a model. The interaction of image forensics and counter-forensics can therefore be framed as *competition for the best image model*.

If counterfeiters succeed in employing superior image models, their attacks are more powerful. Their success does not depend on adjustable definitions of plausibility, but rather on technological weaknesses of forensic algorithms. While at least certain situations allow forensic investigators to lower the level of uncertainty by enforcing higher image quality standards (recall Example 12), it is often more difficult to enter the next round in the competition for the best image model and further advance forensic algorithms.

As has been pointed out in the context of digital watermarking [43, p. 297], robustness is necessary but not sufficient for security. If counter-forensic attacks against a forensic algorithm `decide` with $\texttt{attack} \in \mathscr{P}_{\text{legitimate}}$ exist, this algorithm cannot be considered secure. However, truly secure algorithms need to be reliable under all possible attacks $\texttt{attack} \in \mathscr{P}$.[37]

3.4.3 Adversary Models

A series of influencing factors determines the vulnerability of forensic algorithms to countermeasures, and it is useful to disentangle these effects to understand their individual impact on the security of forensic methods. Traditionally, assumptions about strength and knowledge of the counterfeiter are referred to as *adversary model*, although we have to note that a straightforward adoption from fields like cryptography or other domains of multimedia security is not feasible: digital image forensics does not rely on secret keys. Rather, the decidability of forensic problems solely depends on knowledge about the inherent class likelihoods $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$.

In the following, we discuss what aspects should be considered part of an adversary model in digital image forensics and explore how different settings affect the vulnerability of image forensics algorithms to counter-forensic attacks.

3.4.3.1 Goal of the Attack

Depending on their specific goal, counter-forensic attacks either strive for suppression of identifying traces of the original class $\mathcal{C}$, or synthesis of artificial traces of a particular target class $\mathcal{C}_{\text{target}} \neq \mathcal{C}$ (cf. Section 2.5). The first aim is generally easier to achieve, because no special requirements on the resulting counterfeit exist apart from the general semantic constraint. It is sufficient to move $\boldsymbol{x}^{(\mathcal{C})}$ somewhere outside the decision region $\mathscr{R}_{\mathcal{C}}$. The attack is successful as soon as the forensic investigator cannot trace back the original class anymore. Hence, it is irrelevant to which class the counterfeit is eventually assigned. Counterfeiting a particular target class, on the other hand, is more involved. Here, the counterfeiter needs to address both, the suppression of identifying traces of the original class and the synthesis of artificial traces of the target class.

Example 15 | Sensor Noise Counter-Forensics. In PRNU-based digital camera identification (cf. Example 10), the insertion of the reference noise pattern of a target camera may lead to desired

37 Recent watermarking literature conjectures that robustness and security cannot be achieved at the same time. Rather an acceptable trade-off needs to be found [239, and references therein]. It is an open question whether this also applies to digital image forensics.

decisions for the new class $\mathcal{C}_{\text{target}}$. However, because the (distorted) fingerprint of the true camera is still present, a thorough forensic investigator may find abnormally high likelihood values $\mathcal{P}_{\mathcal{C}}\left(\boldsymbol{x}^{(\tilde{\mathcal{C}}_{\text{target}})}\right)$ suspicious. A more sophisticated attack would thus try to suppress the sensor noise of the original camera before.

3.4.3.2 Knowledge about Function decide

According to Kerckhoffs' principle [120], a system's security must not rely on the secrecy of its algorithms but only on a secret key. Because secret keys do not exist in digital image forensics, a direct application of this widely accepted security paradigm is not possible. Furthermore, Böhme [17] pointed out that a general hurdle to the adoption of Kerckhoffs' principle to empirical disciplines lies in the incognizability of distributions $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$. A strict interpretation would require to grant counterfeiters super-natural capabilities. With reference to Kerckhoffs' principle, it still seems reasonable to assume that counterfeiters have full knowledge of the forensic investigator's image model and decision rule, respectively, and to accept that specific security properties only hold valid as long as counterfeiters have not found better models.

3.4.3.3 Image Model of the Attacker

The better assumptions about $\mathcal{P}_{\mathcal{C}'_0}$ and $\mathcal{P}_{\mathcal{C}'_1}$ adhere to reality the better can counterfeiters exploit shortcomings of existing image models. As such, the image model of the attacker is probably the most crucial part of any adversary model for image forensics, and its sophistication is directly proportional to the security of algorithm `decide`. As pointed out in Section 3.4.2, epistemic bounds dictate that no conclusiveness exists because forensic investigators will always have the chance to make improvements to their own model.

3.4.3.4 Access to Function generate

Image models reflect the general knowledge that both forensic investigators and counterfeiters have of instances of function `generate` and the corresponding class likelihoods. The security of forensic algorithms further depends on the capability to sample outcomes of relevant instances of function `generate`. It is well possible that a counterfeiter is in possession of a superior image model in theory, but cannot put it in use satisfactory because of missing access to functions $(\texttt{acquire}, \texttt{process}) \in (\mathscr{A} \times \mathscr{P})^{(\mathcal{C}'_0)}$. Consider for instance the above Example 15 of inserting a new PRNU fingerprint into an image to pretend a different source digital camera. While efficient methods for this purpose undoubtedly exist, they require access to the target digital camera to prevent that residues of publicly available images, which have been used to create the fingerprint, remain detectable in the counterfeit [92].

3.4.3.5 Admissible Semantic Resolution

The capability to distinguish between authentic and inauthentic images further depends on the images' available semantic resolution (cf. Sections 3.1.2 and 3.3.3.2). If forensic investigators are in the position to impose strict requirements on the minimum necessary image quality, they can effectively limit the design space of counter-forensic attacks by forcing counterfeiters to come up with image models of ever-increasing sophistication. In this view, the *admissible*

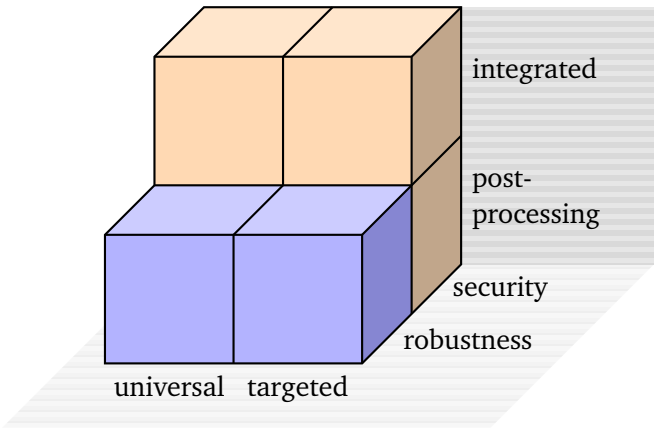


Figure 3.8: Design space for counter-forensic techniques.

semantic resolution is a type of security parameter. Similar to the key length in cryptography or the inverse embedding rate in steganography, it controls the attacker's probability of success.

3.4.3.6 Computational Bounds

Realistic forensic analyses will only yield reliable results against the backdrop of evaluating a large number of different class likelihoods for a even larger number of images.[38] Combined with the high dimensionality of typical digital images, this calls for considering also the computational resources of forensic investigators and counterfeiters as building blocks to an adversary model. Computational power can, within certain limits, impact the quality of available image models in two directions. First, unconstrained access to computational resources allows (in theory) to handle image models of arbitrary complexity and prevents from the use of sub-optimal procedures. As a typical example, consider the test of a particular PRNU signal against a large database of known digital camera fingerprints, where it has been proposed to rely on trimmed [90] or aggregated [13] signal representations to keep the problem computationally tractable. On a second dimension, computational power can certainly mitigate the problem of sampling images, which ultimately increases the information available about distributions $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$. Still, even computationally unconstrained forensic investigators and/or counterfeiters are not free of epistemic bounds. More computations cannot compensate for inherently incomplete knowledge about the incognizable reality (and vice versa) [18, p. 101].

3.4.4 Classification of Counter-Forensic Techniques

A first classification of counter-forensic techniques was already discussed in Section 3.4.2, where we distinguished between attacks based on robustness and security properties of forensic algorithms. In the following we present two further dimensions along which counter-forensics can be classified. While *integrated* and *post-processing* attacks vary in their position in the image generation process, *targeted* and *universal* attacks differ in the (range of) attacked forensic algorithms. Figure 3.8 illustrates our classification and the following subsections discuss each of the dimensions in more detail.

38 A large class space is of particular relevance in source device identification. The analysis of a considerable amount of images is always necessary to infer or validate reasonable image models.

3.4.4.1 Integrated and Post-Processing Attacks

Post-processing attacks modify images $\boldsymbol{x}^{(k)}$ such that the resulting counterfeits $\boldsymbol{x}^{(\tilde{l})}$ do not exhibit traces of the original class $\mathcal{C}_k$ anymore (cf. Example 13). Figure 3.9 illustrates that such attacks can be thought of as additional processing step $\texttt{attack} \in \mathscr{P}$, which supplements the original generation process. They are hence implemented as direct application of Equation (3.36) and may take advantage of robustness issues or security weaknesses.

Integrated attacks, on the other hand, interact with or replace parts of the original image generation process. Instead of $\boldsymbol{x}^{(k)}$, the counterfeit $\boldsymbol{x}^{(\tilde{l})}$ is generated directly by a tuple $(\texttt{acquire}', \texttt{process}')$. The modified functions $\texttt{acquire}'$ and $\texttt{process}'$ are specifically designed to avoid the formation of identifying traces of the original class or to mimic characteristics of the target class (see also Figure 3.9). In the aforementioned Example 13, an integrated attack would directly transform the authentic image $\boldsymbol{x}_1^{(1)}$ to a semantically different counterfeit $\boldsymbol{x}_3^{(\tilde{1})}$ without ever releasing the detectable manipulation $\boldsymbol{x}_2^{(0)}$. We note that this procedure is also covered by our formal description of counter-forensic attacks in Definition 13, because it is always possible to express the (imaginary) map $\boldsymbol{x}_2^{(0)} \mapsto \boldsymbol{x}_3^{(\tilde{1})}$ in terms of a post-processing function $\texttt{attack}$. As integrated methods obviously require deep knowledge of the image generation process, they do not address robustness issues of forensic algorithms by definition. This is also indicated in Figure 3.8, where the corresponding regions are left blank.

Integrated methods are mainly relevant for manipulation detectors. Here, counterfeiters are hardly restricted in the choice of image processing primitives and can replace undesirable components according to their needs. Integrated attacks to impede source identification are less obvious.[39] Nevertheless, it is conceivable to modify software for raw image processing for counter-forensic purposes. With freely available open-source firmware modifications, device-internal counter-forensics may become a serious threat to forensic investigators since essential device-specific traces need not leave the device at all, cf. Section 3.3.3.3.

3.4.4.2 Targeted and Universal Attacks

A further classification is borrowed from the context of steganalysis [66] and digital watermarking [43]. We call an attack *targeted*, if it exploits particulars and weaknesses of one specific forensic algorithm $\texttt{decide}$, which the counterfeiter usually knows. Such vulnerabilities directly relate to the image model implemented in the forensic algorithm. Clearly, it is possible (and likely) that other forensic algorithms using alternative or improved image models can detect such counterfeits.

Conversely, *universal attacks* try to maintain or correct as many statistical properties of the image in order to conceal manipulations even when presented to unknown forensic tools. This is by far the more difficult task, and it is an open research question whether image models can be found good enough to sustain analysis with combinations of forensic algorithms. Even if we assume that targeted attacks against all known forensic algorithms exist, a simple combination of them in general will not lead to a universal attack—at least when typical low-dimensional image models are employed, which ignore the interference and interdependence of different attacks.

39 Pathological attacks can be constructed by capturing a scene with a completely different device.

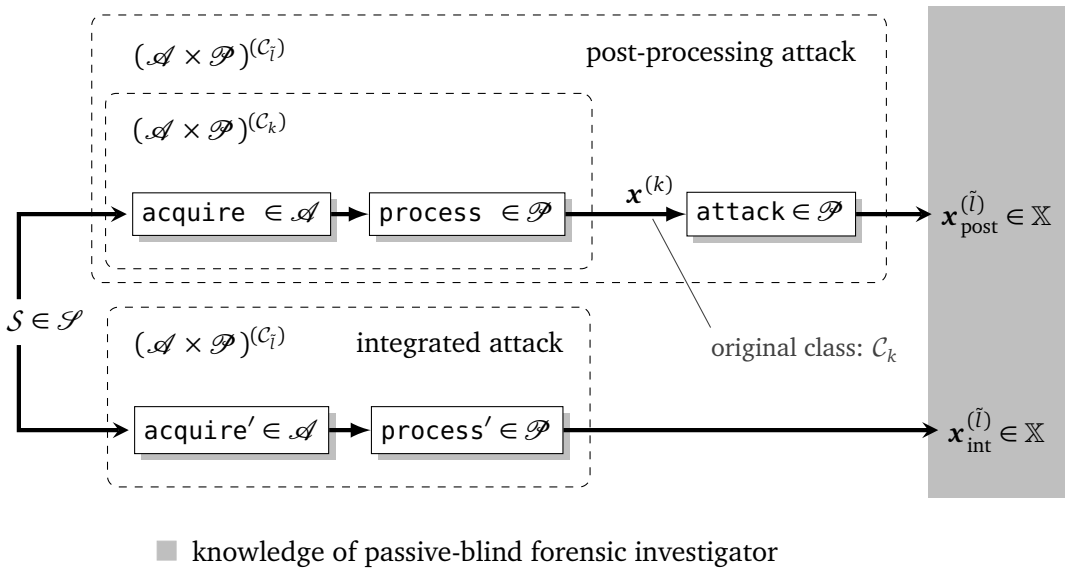


Figure 3.9: Post-processing and integrated counter-forensic attacks. Post-processing attacks suppress and/or synthesize identifying traces subsequent to the original image generation process. Integrated attacks directly replace the original process with a counter-forensic variant.

A weaker, yet more practical form of universal attacks exploits poor robustness properties and uses lossy but legitimate processing whenever plausible. Given most forensic algorithms' strong dependence on the available semantic resolution, this type of attack will often be the most convenient way to deliberately mislead a multitude of forensic schemes at the same time. Recall that this variant of universal attacks is always a trade-off between originality and plausibility of the counterfeit. Even if strong quantization removes identifying traces of the original class, it very likely precludes claims about the originality of the corresponding image (cf. Section 3.3.3.2).

3.5 Relations to Prior Work

The previous sections introduced and discussed a formal framework for the description of digital image forensics and counter-forensic attacks. As we have mentioned in the very beginning, reflections on theoretical underpinnings of this emerging research field have been mostly subordinated to endeavors of advancing specific detection techniques. Nevertheless, our classification framework is clearly driven by the vast number of practical schemes that, in various ways, make use of long-established statistical classification and pattern recognition techniques. Already one of the very first works on digital image forensics alludes to classification methods and likelihood ratio tests [79]. An explicit mention of the relation between class space and image generation process first appeared in 2004 [7, 125, 196]. Yet also many of our more specific observations with regard to image authenticity, image models, robustness and security have appeared in the existing body of literature, although mostly in implicit form. For instance, both Popescu [194, p. 5] and Wang et al. [231, p. 319] hint to potential issues that may arise from legitimate post-processing. Robustness of forensic algorithms (mainly to JPEG

compression) has in general received some interest throughout the literature. Farid and Lyu [60] as well as Lukáš et al. [153] were among the first to mention that 'counter-attacks' and 'malicious post-processing' against specific forensic algorithms are a potential threat to forensic investigators.[40] Implicit adversary models appeared early on in the work of Farid et al., where it was usually assumed that sophisticated targeted attacks are beyond the skills of average image forgers [198, i. a.]. Goljan et al. [91] discuss specific 'attack scenarios' and reflected on strength and sophistication of potential counterfeiters. Huang [106] and Ng [173] first indicated that counterfeiters are subject to a semantic constraint.[41]

Only few works have reasoned about the theoretical background of a broader image forensics and counter-forensics framework. In the following, we discuss connections of our work to earlier definitions of authenticity (Section 3.5.1) and component forensics (Section 3.5.2).

3.5.1 Image Authenticity

Ng [173, p. 6] defines authenticity as the property of digital images to «*represent a witness to an actual event, place, or time*». He further mentions two factors that control authenticity, namely the *natural scene quality* and the *imaging process quality* of digital images. The former relates to the «*physics of the real-world light transport*» (i. e., the natural phenomenon, which is projected to a digital image, see also Section 2.1.4) and the latter to characteristics of the image acquisition device [181, p. 387]. The rationale here is that authentic images are valid representations of natural phenomena (natural scene quality), which exhibit identifying traces of particular image acquisition devices (imaging process quality). Both quantities can be examined to infer the authenticity of an image under investigation, which can be either scene-authentic, imaging-process-authentic, or both [177]. The definition of natural scene quality acknowledges both the empirical nature and the high dimensionality of the problem. Consequently, only low-dimensional image models are available, which should reflect natural image statistics, device parameters as well as scene constraints [181, pp. 400–401]. Although rather informal, Ng's setup clearly contains most ingredients to our idea of authenticity, and it is straightforward to plug Definitions 4 and 5 into his framework to obtain the first part of Definition 7: all original natural images are authentic. Yet, because inauthenticity is treated only implicitly via the absence of authenticating characteristics, no direct parallels to the notions of semantic equivalence and legitimate post-processing exist.

3.5.2 Component Forensics and Classifiability

The decidability of a forensic problem hinges on the notion of identifying traces of particular parameter spaces $\Theta^{(C)}$. Distinguishing between outputs of different image generation functions (i. e., between samples drawn from the associated class likelihoods) ultimately means to infer (a subset of) the corresponding parameters $\boldsymbol{\theta}$. Swaminathan et al. [218, 220, 224] assume that each component of an acquisition device is characterized by a set of parameters $\boldsymbol{\theta}_c \subset \boldsymbol{\theta}$, where $\boldsymbol{\theta}_c$ denotes the parameters of the c-th component. The goal of *component forensics*

40 For completeness, we note that Huang's paper on the synthesis of color filter array artifacts [106] was the first publication that explicitly addressed counter-forensics. Unfortunately, many of the definitions and concepts therein remain unclear. In particular, Huang made the definition of the true class of an image depend on partitions of the image space, which gives rise to decision ambiguities (cf. our discussion preparatory to Definition 9).

41 However, both authors do no explicitly link their rather vague requirements, namely $\boldsymbol{x}^{(\tilde{1})} \approx \boldsymbol{x}^{(0)}$ [106] and a minimal visual quality loss between $\boldsymbol{x}^{(0)}$ and $\boldsymbol{x}^{(\tilde{1})}$ [173, p. 117], to the semantic meaning of image $\boldsymbol{x}^{(0)}$.

is then to assign a given image under analysis to one out of K_c possible configurations $\boldsymbol{\theta}_c^{(k)}$, $0 \leq k < K_c$, a component may take [222]. Because passive–blind investigators cannot access individual components but have to examine the device as a black box, component forensics of a N-component device here boils down to inference about a set of parameters $\{\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_{N-1}\}$. By defining classes $\mathcal{C}$ to encapsulate specific parameter combinations

$$\Theta^{(\mathcal{C})} = \left\{ \boldsymbol{\theta} \,\middle|\, \boldsymbol{\theta}_0^{(k_0^{\mathcal{C}})}, \ldots, \boldsymbol{\theta}_{N-1}^{(k_{N-1}^{\mathcal{C}})} \in \boldsymbol{\theta} \right\}, \quad 0 \leq k_c^{\mathcal{C}} < K_c \,,$$

it is clear that component forensics is a special image forensics problem, which is defined over a class space of size $K = |\mathscr{C}| = \prod_{c=0}^{N-1} K_c$, cf. Definition 8.

To measure the decidability of the corresponding forensic problem, Swaminathan et al. [222] introduce the notion of 'classifiability'. More specifically, a class space $\mathscr{C} = \{\mathcal{C}_0, \ldots, \mathcal{C}_{K-1}\}$ is said to be (non-intrusively) classifiable, if for each $\mathcal{C}_k \in \mathscr{C}$:

$$\forall \boldsymbol{x} \in \mathbb{X} \qquad \Pr(\mathcal{C}_k \,|\, \boldsymbol{x}) \geq \Pr(\mathcal{C}_l \,|\, \boldsymbol{x}) \quad \forall l \neq k \,, \text{ and} \tag{3.39}$$

$$\exists \boldsymbol{x}' \in \mathbb{X} \qquad \Pr(\mathcal{C}_k \,|\, \boldsymbol{x}') > \Pr(\mathcal{C}_l \,|\, \boldsymbol{x}') \quad \forall l \neq k \,. \tag{3.40}$$

We remark that this definition is of limited use in typical passive–blind settings, where the forensic investigator has no influence on which images to analyze. A K-class problem will be considered classifiable even if the forensic investigator can do no better than random guessing in the vast majority of $|\mathbb{X}| - K$ cases, $K \lll |\mathbb{X}|$.[42] We therefore believe that any binary measure of passive–blind decidability should really be a measure of *un*decidability in order to avoid misleading conclusions.[43] Another option is of course to quantify (un)decidability, for instance using Kullback–Leibler divergence (cf. Definition 11) or alternative multi-distribution divergence measures such as Jensen-Shannon divergence [145].

3.6 Relations to Steganography and Digital Watermarking

We close this chapter by broadening the perspective a bit further. Counter-forensics and its competition with digital image forensics should not be considered an isolated research field. It should rather be seen in close relation to established disciplines in the broader field of information hiding and multimedia security. This way, obvious parallels can help to foster and deepen the understanding of research questions in digital image forensics and counter-forensics that may have already been identified and discussed elsewhere.

In particular, the general definitions in Sections 3.1 to 3.4, namely

- ▹ image generation processes that introduce identifying traces and thus span a class space,
- ▹ intentional class changes that are subject to constraints, and
- ▹ decision rules to infer the unknown class of an image

42 For a binary class space and uniform priors, any two distributions $\mathcal{P}_0(\boldsymbol{x})$ and $\mathcal{P}_1(\boldsymbol{x})$ with $\mathsf{D}_{\mathrm{KL}}(\mathcal{P}_0, \mathcal{P}_1) = \epsilon > 0$, will be called classifiable, independent of how small ϵ is.

43 A somewhat different situation arises in non-blind settings *if* the forensic investigator can choose which images to analyze. Swaminathan et al. [218] name the investigation of patent infringement as a particular example.

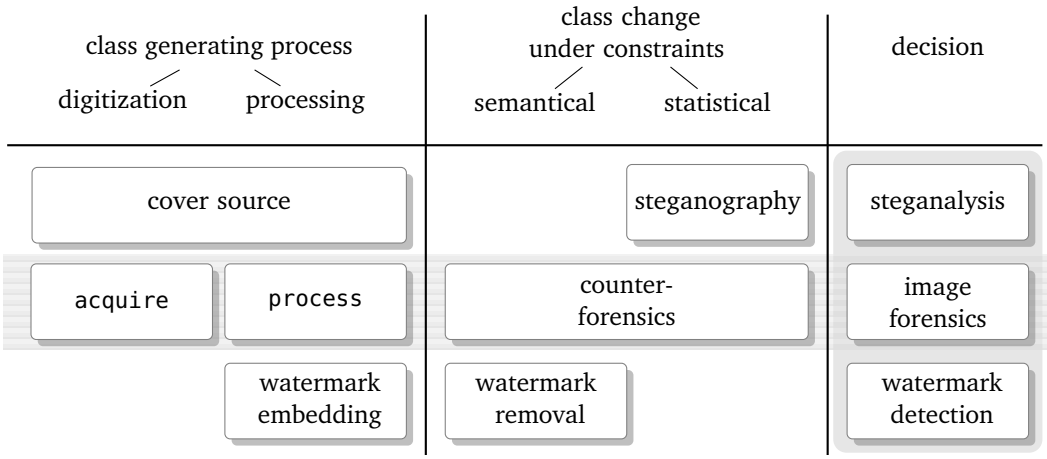


Figure 3.10: Relation of image forensics and counter-forensics to other fields in information hiding.

have already appeared in very similar form in the context of steganographic communication [66, 18] and digital watermarking [43], and their specific notions of security. Figure 3.10 illustrates how these disciplines relate to digital image forensics and counter-forensics. It serves as a blueprint for the following discussion. The process chain from image generation, via class change, to decision is depicted in columns from left to right. Important distinctions, such as between empirical acquisition and deterministic processing or the nature of constrains are also reflected. Focal areas of the specific sub-fields of information hiding are arranged in rows so that corresponding hiding, detection, and deception techniques can be associated horizontally. Similarities of digital image forensics and counter-forensics with related tasks in steganographic communication and digital watermarking appear as vertical relations.

Counter-forensics shares common goals with *steganography*. By embedding a secret message, a cover image's class is changed from $\mathcal{C}_1$ to the class of stego images $\mathcal{C}_0$. Both steganography and counter-forensics try to hide the very fact of a class change, and their success can be measured by the Kullback–Leibler divergence between the two conditional probability distributions $\mathcal{P}_{\mathcal{C}_1}$ and $\mathcal{P}_{\mathcal{C}_0}$ (cf. Equation (3.19)). This imposes statistical constraints on both. Steganography differs from counter-forensics in the amount and source of information to hide. Most steganographic algorithms are designed to embed a message by minimizing distortion, thereby preserving the cover's semantic meaning. Counter-forensics, by contrast, conceals the mere information that larger parts of the original image $\boldsymbol{x}^{(1)}$ have been modified, often with the aim to change its semantic meaning. The new semantic meaning of the counterfeit $\boldsymbol{x}^{(0)}$ can be regarded as the 'message' to be transmitted.

Steganalysis, as a counterpart to steganography, aims at unveiling the presence of a hidden message in a specific image without having access to the original cover. A general analogy between steganalysis and image forensics becomes evident if we consider the act of counterfeiting as information which is hidden inconspicuously in an image. This suggests that counter-forensics—similar to steganography, where capacity and security are considered as competing design goals—needs to trade off the amount of information to hide and detectability. The stronger a manipulating operation interferes with the inherent image structure, the harder it is to feign an authentic image.

Another analogy exists between counter-forensics and *attacks against robust digital watermarking* schemes, where images of class $\mathcal{C}_0$ are generated by a *watermark embedding* process. In contrast to steganalysis, attacks against (robust) digital watermarks are designed to merely remove the embedded information, i. e., changing a watermarked image's class to $\mathcal{C}_1$, while retaining the semantic meaning and to some degree the perceptual quality of the cover. In this sense, identifying traces in digital image forensics can be understood as inherent watermark, which counter-forensics aim to suppress or change. Since attacks against digital watermarks do not specifically address statistical undetectability, the robustness issues of digital watermarking schemes may find a correspondent in digital image forensics. Similar to digital watermarking [114], forensic investigators also wish that the performance of their algorithms degrades gently as a function of image quality loss.

Note that the above parallels, on a technical level, correspond to the two different approaches to counter-forensics, as discussed in Section 3.4.4.1. Integrated attacks are more closely related to steganography (hiding traces of a class change by design) whereas post-processing attacks have stronger similarities to attacks against digital watermarking (remove identifying traces).

We conclude this excursion to related fields with a note of caution. The rather informal description of the above similarities should not be read as a suggestion to disregard important particularities of each field. Although all discussed disciplines are in general tied to epistemic bounds (as long as empirical covers are involved [17]), the primary source of differences lies in the role of the class generating process. More specifically, digital watermarks can be *designed* to have certain robustness or security properties, i. e., type and strength of identifying traces are under full control of the watermark designer and can be shaped according to particular requirements (cf. Section 2.2.1). Also in steganography, the class generating process is more or less under control of the steganographer. Cover images are solely a means to communicate hidden messages. This leaves more degrees of freedom to choose the cover and the strength of distortion due to embedding. This is totally different in digital image forensics and counter-forensics. Neither can forensic investigators control the quality of identifying traces (apart from imposing a certain admissible semantic resolution, cf. Section 3.4.3.5), nor can counterfeiters overcome their semantic constraints. In the long run, their only option is to constantly gather knowledge about image generation processes and to eventually improve their image models.

3.7 Summary

This chapter has devised a formal framework of digital image forensics and counter-forensics. With regard to the underpinnings of our framework, we see the following main contributions of this chapter in

- the formalization of a *universal image generation process* as a proxy for arbitrary image acquisition and post-processing procedures (Section 3.1.1), which allowed us to give
- a formal definition of relevant attributes (most importantly the notion of *authenticity*) of digital images in the context of digital image forensics (Sections 3.1.1.2 and 3.1.2).

Based on this image generation process, we then formulated image forensics as a classification problem and discussed its decision-theoretic foundations. The viability of this perspective on digital image forensics has already been confirmed before by numerous concrete instances

of practical forensic algorithms in the body of literature. We adapted the underlying ideas and concepts to build a coherent framework that is not limited to specific sub-problems (e. g., source identification or detection of post-processing) but rather covers digital image forensics in a very broad sense. Particular emphasis was put on

- the exploration of *fundamental limits* of digital image forensics, which are imposed by the decision-theoretic framework (Section 3.2.2.2) and by practical constraints (Section 3.3).

We interpreted this framework in a security context and gave

- a formal definition of *counter-forensics* fully consistent with the previously introduced notations (Section 3.4.1), accompanied by
- a discussion of general *robustness and security* properties of forensic algorithms (Section 3.4.2) and a *classification of counter-forensic attacks* (Section 3.4.4).

Parallels to related disciplines like steganography or digital watermarking indicate that the notion of a universal image generation process and the resulting decision problem is not specific to digital image forensics.

Recent works in the literature emphasize that only slightly more sophisticated image models—on the side of either the forensic investigator or the counterfeiter—can be sufficient to devise more reliable detectors and/or attacks. In general, we expect forensic algorithms to remain under constant development. At the same time, a growing number of publications on counter-forensics suggests an increased awareness of security-related questions in the field of digital image forensics. First works addressing the problem of "countering counter-forensics" [92, 139, 228] indicate that the competition for superior image models has just begun.

Improvements on the robustness of forensic algorithms are of equally high practical relevance. Digital images are often downsampled and/or JPEG compressed before being made public. While being perfectly plausible in many situations, low-resolution images are likely to conceal a considerable part of the image generation process' identifying traces from forensic algorithms. As for JPEG images, the literature also provides an ample body of tools to exploit specific characteristics of the compression pipeline, for instance, to detect repeated saving in the JPEG format [196, 31, 188]. Yet no methods to robustly deal with combinations of double compression *and* downsampling are presently known. With respect to strong downsampling, we generally remain more pessimistic, at least as far as detectors based on periodic interpolation artifacts are concerned. Although recent frequency domain approaches discuss first promising alternatives [61], it is too early to prognosticate whether forensic investigators can ever expect a moderate level of robustness, at least.

In conclusion, both security and robustness of forensic algorithms will equally have to be addressed in future works. In this sense, the future development of digital image forensics is likely to resemble a process similar to the evolvement of related information hiding disciplines, where a fruitful cat-and-mouse game has led to hiding and detection algorithms of considerable sophistication [73, 189, 135, i .a.]. In general, we believe that a lively and mutual interaction will prove beneficial to the whole field of information hiding. Not only can the rather young field of digital image forensics and counter-forensics learn from the more established branches, also steganography and steganalysis, which both have to cope with heterogenous image sources, can gain from findings in digital image forensics to conceive better, or at least adaptive, image models [18, 9]. Moreover, the literature now reports digital watermarking schemes that directly

interact with the image generation process [165], which suggests to employ (counter-)forensic techniques as building blocks for the design of attacks against digital watermarks and the detection thereof. Yet the strong overlap not only calls for interaction on the practical side. Formalizing unified theoretical foundations seems a further promising goal to advance the overall field.

Bibliography

The citation policy is as follows: if journal papers extend earlier conference or workshop manuscripts by the same authors (and the time of publication is not critical), we refer only to the respective journal version.

[1] Gazi N. Ali, Aravind K. Mikkilineni, Pei-Ju Chiang, Jan P. Allebach, George T.-C. Chiu, and Edward J. Delp. "Intrinsic and Extrinsic Signatures for Information Hiding and Secure Printing with Electrophotographic Devices". In: *NIP 19: International Conference on Digital Printing Technologies* (New Orleans, LA, Sept. 28–Oct. 3, 2003). Bellingham, WA: SPIE Press, 2003, pp. 511–515.
http://cobweb.ecn.purdue.edu/~prints/public/papers/nip03-ali.pdf.

[2] Gazi N. Ali, Pei-Ju Chiang, Aravind K. Mikkilineni, George T.-C. Chiu, Edward J. Delp, and Jan P. Allebach. "Application of Principal Components Analysis and Gaussian Mixture Models to Printer Identification". In: *NIP 20: International Conference on Digital Printing Technologies* (Salt Lake City, UT, Oct. 31–Nov. 5, 2004). Bellingham, WA: SPIE Press, 2004, pp. 301–305.
http://cobweb.ecn.purdue.edu/~prints/public/papers/nip04-ali.pdf.

[3] Erwin J. Alles, Zeno J. Geradts, and Cor. J. Veenman. "Source Camera Identification for Heavily JPEG Compressed Low Resolution Still Images". In: *Journal of Forensic Sciences* 54.3 (May 2009), pp. 628–638.
doi:10.1111/j.1556-4029.2009.01029.x.

[4] Paul Alvarez. "Using Extended File Information (EXIF) File Headers in Digital Evidence Analysis". In: *International Journal of Digital Evidence* 2.3 (Winter 2004).
http://www.utica.edu/academic/institutes/ecii/publications/articles/A0B1F944-FF4E-4788-E75541A7418DAE24.pdf.

[5] Irene Amerini, Lamberto Ballan, Roberto Caldelli, Alberto Del Bimbo, and Giuseppe Serra. "A SIFT-based Forensic Method for Copy-Move Attack Detection and Transformation Recovery". In: *IEEE Transactions on Information Forensics and Security* 6.3 (Sept. 2011), pp. 1099–1110.
doi:10.1109/TIFS.2011.2129512.

[6] Vassilis Athitsos, Michael J. Swain, and Charles Frankel. "Distinguishing Photographs and Graphics on the World Wide Web". In: *IEEE Workshop on Content-Based Access of Image and Video Libraries, CBAIVL '97* (San Juan, Puerto Rico, June 20, 1997). 1997, pp. 10–17.
doi:10.1109/IVL.1997.629715.

[7] İsmail Avcibaş, Sevinç Bayram, Nasir Memon, Mahalingam Ramkumar, and Bulent Sankur. “A Classifier Design For Detecting Image Manipulations”. In: *International Conference on Image Processing, ICIP 2004* (Singapore, Oct. 24–27, 2004). Vol. 4. 2004, pp. 2645–2648.
doi:10.1109/ICIP.2004.1421647.

[8] Shai Avidan and Ariel Shamir. “Seam Carving for Content-Aware Image Resizing”. In: *ACM Transactions on Graphics* 26.3 (July 2007): *Proceedings of ACM SIGGRAPH 2007*.
doi:10.1145/1276377.1276390.

[9] Mauro Barni, Giacomo Cancelli, and Annalisa Esposito. “Forensics Aided Steganalysis of Heterogeneous Images”. In: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2010* (Dallas, TX, Mar. 14–19, 2010). 2010, pp. 1690–1693.
doi:10.1109/ICASSP.2010.5495494.

[10] B. E. Bayer. *Color Imaging Array*. US Patent, 3 971 065. 1976.

[11] Sevinç Bayram, Husrev T. Sencar, and Nasir Memon. “Improvements on Source Camera-Model Identification Based on CFA”. In: *Advances in Digital Forensics II, IFIP International Conference on Digital Forensics* (Orlando, FL, Jan. 29–Feb. 1, 2006). Ed. by Martin S. Olivier and Sujeet Shenoi. Vol. 222. IFIP Advances in Information and Communication Technology. Boston, MA: Springer, 2006, chap. 23, pp. 289–299.
doi:10.1007/978-0-387-36891-7_23.

[12] Sevinç Bayram, Husrev T. Sencar, and Nasir Memon. “Classification of Digital Camera-Models Based on Demosaicing Artifacts”. In: *Digital Investigation* 5.1–2 (Sept. 2008), pp. 46–59.
doi:10.1016/j.diin.2008.06.004.

[13] Sevinç Bayram, Husrev T. Sencar, and Nasir Memon. “Efficient Techniques For Sensor Fingerprint Matching In Large Image and Video Databases”. In: *Media Forensics and Security II* (San Jose, CA, Jan. 18–20, 2010). Ed. by Nasir D. Memon, Jana Dittmann, Adnan M. Alattar, and Edward J. Delp. Vol. 7541. Proceedings of SPIE. Bellingham, WA: SPIE, 2010, 754109.
doi:10.1117/12.845737.

[14] Christoper M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[15] Greg J. Bloy. “Blind Camera Fingerprinting and Image Clustering”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.3 (Mar. 2008), pp. 532–535.
doi:10.1109/TPAMI.2007.1183.

[16] Paul Blythe and Jessica Fridrich. “Secure Digital Camera”. In: *Digital Forensic Research Workshop* (Baltimore, MD, Aug. 11–13, 2004). 2004.
http://www.dfrws.org/2004/day3/D3-Blyth_Secure_Digital_Camera.pdf.

[17] Rainer Böhme. “An Epistemological Approach to Steganography”. In: *Information Hiding, 11th International Workshop, IH 2009. Revised Selected Papers* (Darmstadt, Germany, June 8–10, 2009). Ed. by Stefan Katzenbeisser and Ahmad-Reza Sadeghi. Vol. 5806. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 15–30.

doi:10.1007/978-3-642-04431-1_2.

[18] Rainer Böhme. *Advanced Statistical Steganalysis*. Berlin, Heidelberg: Springer-Verlag, 2010.
doi:10.1007/978-3-642-14313-7.

[19] Rainer Böhme and Matthias Kirchner. "Counter-Forensics: Attacking Image Forensics". In: *Digital Image Forensics*. Ed. by Husrev Taha Sencar and Nasir Memon. Springer-Verlag, 2012, pp. 327–366.

[20] Dino A. Brugioni. *Photo Fakery. The History and Techniques of Photographic Deception and Manipulation*. Dulles, VA: Brassey's, 1999.

[21] Robert W. Buccigrossi and Eero P. Simoncelli. "Image Compression via Joint Statistical Characterization in the Wavelet Domain". In: *IEEE Transactions on Image Processing* 8.12 (Dec. 1999), pp. 1688–1701.
doi:10.1109/83.806616.

[22] Christian Cachin. "An Information-Theoretic Model for Steganography". In: *Information and Computation* 192.1 (July 2004), pp. 41–56.
doi:10.1016/j.ic.2004.02.003.

[23] Roberto Caldelli, Irene Amerini, and Francesco Picchioni. "A DFT-Based Analysis to Discern Between Camera and Scanned Images". In: *International Journal of Digital Crime and Forensics* 2.1 (2010), pp. 21–29.
doi:10.4018/jdcf.2010010102.

[24] Gang Cao, Yao Zhao, Rongrong Ni, and Huawei Tian. "Anti-forensics of Contrast Enhancement in Digital Images". In: *MM&Sec'10, Proceedings of the 2010 ACM SIGMM Multimedia & Security Workshop* (Rome, Italy, Sept. 9–10, 2010). New York: ACM Press, 2010, pp. 25–34.
doi:10.1145/1854229.1854237.

[25] Hong Cao and Alex C. Kot. "Accurate Detection of Demosaicing Regularity for Digital Image Forensics". In: *IEEE Transactions on Information Forensics and Security* 4.4 (Dec. 2009), pp. 899–910.
doi:10.1109/TIFS.2009.2033749.

[26] Hong Cao and Alex C. Kot. "Identification of recaptured photographs on LCD screens". In: *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2010* (Dallas, TX, Mar. 14–19, 2010). 2010, pp. 1790–1793.
doi:10.1109/ICASSP.2010.5495419.

[27] Brian Carrier. "Defining Digital Forensic Examination and Analysis Tools Using Abstraction Layers". In: *International Journal of Digital Evidence* 1.4 (Winter 2003).
http://www.utica.edu/academic/institutes/ecii/publications/articles/A04C3F91-AFBB-FC13-4A2E0F13203BA980.pdf.

[28] Oya Çeliktutan, Bülent Sankur, and İsmail Avcibaş. "Blind Identification of Source Cell-Phone Model". In: *IEEE Transactions on Information Forensics and Security* 3.3 (Sept. 2008), pp. 553–566.
doi:10.1109/TIFS.2008.926993.

[29] Edward Chang, Shiufun Cheung, and Davis Y. Pan. “Color filter array recovery using a threshold-based variable number of gradients”. In: *Sensors, Cameras, and Applications for Digital Photography* (San Jose, CA, Jan. 27, 1999). Ed. by Nitin Sampat and Thomas Yeh. Vol. 3650. Proceedings of SPIE. Bellingham, WA: SPIE, 1999, pp. 36–43. doi:10.1117/12.342861.

[30] David L. Chaum. “Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms”. In: *Communications of the ACM* 24.2 (Feb. 1981), pp. 84–90. doi:10.1145/358549.358563.

[31] Chunhua Chen, Yun Q. Shi, and Wei Su. “A Machine Learning Based Scheme for Double JPEG Compression Detection”. In: *19th International Conference on Pattern Recognition, ICPR 2008* (Tampa, FL, Dec. 8–11, 2008). 2008. doi:10.1109/ICPR.2008.4761645.

[32] Mo Chen, Jessica Fridrich, Miroslav Goljan, and Jan Lukáš. “Determining Image Origin and Integrity Using Sensor Noise”. In: *IEEE Transactions on Information Forensics and Security* 3.1 (Mar. 2008), pp. 74–90. doi:10.1109/TIFS.2007.916285.

[33] Sz-Han Chen and Chiou-Ting Hsu. “Source Camera Identification Based on Camera Gain Histogram”. In: *IEEE International Conference on Image Processing, ICIP 2007* (San Antonio, TX, Sept. 16–19, 2007). Vol. 4. 2007, pp. IV–429–432. doi:10.1109/ICIP.2007.4380046.

[34] Yi-Lei Chen and Chiou-Ting Hsu. “Detecting Recompression of JPEG Images via Periodicity Analysis of Compression Artifacts for Tampering Detection”. In: *IEEE Transactions on Information Forensics and Security* 6.2 (June 2011), pp. 396–406. doi:10.1109/TIFS.2011.2106121.

[35] Pei-Ju Chiang, Nitin Khanna, Aravind K. Mikkilineni, Maria V. Oritz Segovia, Jan P. Allebach, George T.-C. Chiu, and Edward J. Delp. “Printer and Scanner Forensics: Models and Methods”. In: *Intelligent Multimedia Analysis for Security Applications*. Ed. by Husrev T. Sencar, Sergio Velastin, Nikolaos Nikolaidis, and Shiguo Lian. Studies in Computational Intelligence 282. Berlin, Heidelberg: Springer Verlag, 2010, pp. 145–187. doi:10.1007/978-3-642-11756-5_7.

[36] Chinese Academy of Sciences, Institute of Automation. *CASIA Tampered Image Detection Evaluation Database*. 2009/2010. http://forensics.idealtest.org.

[37] Kai San Choi, Edmund Y. Lam, and Kenneth K. Y. Wong. “Automatic Source Camera Identification Using the Intrinsic Lens Radial Distortion”. In: *Optics Express* 14.24 (Nov. 2006), pp. 11551–11565. doi:10.1364/OE.14.011551.

[38] C. K. Chow. “An Optimum Character Recognition System Using Decision Functions”. In: *IRE Transactions on Electronic Computers* 6.4 (Dec. 1957), pp. 247–254. doi:10.1109/TEC.1957.5222035.

[39] Vincent Christlein, Christian Riess, and Elli Angelopoulou. “A Study on Features for the Detection of Copy-Move Forgeries”. In: *Sicherheit 2010. Konferenzband der 5. Jahrestagung des Fachbereichs Sicherheit der Gesellschaft für Informatik e. V. (GI)* (Berlin, Germany, Oct. 5–7, 2010). Ed. by Felix C. Freiling. Bonn: Gesellschaft für Informatik e. V., 2010, pp. 105–116.
http://subs.emis.de/LNI/Proceedings/Proceedings170/P-170.pdf.

[40] Vincent Christlein, Christian Riess, and Elli Angelopoulou. “On Rotation Invariance in Copy-Move Forgery Detection”. In: *IEEE International Workshop on Information Forensics and Security, WIFS 2010* (Seattle, WA, Dec. 12–15, 2010). 2010.
doi:10.1109/WIFS.2010.5711472.

[41] Corinna Cortes and Vladimir Vapnik. “Support-Vector Networks”. In: *Machine Learning* 20.3 (Sept. 1995), pp. 273–297.
doi:10.1007/BF00994018.

[42] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc., 2006.

[43] Ingemar J. Cox, Matthew L. Miller, Jeffrey A. Bloom, Jessica Fridrich, and Ton Kalker. *Digital Watermarking and Steganography*. Morgan Kaufmann, 2008.

[44] Douglas W. Cromey. “Avoiding Twisted Pixels: Ethical Guidelines for the Appropriate Use and Manipulation of Scientific Digital Images”. In: *Science and Engineering Ethics* 16.4 (Dec. 2010), pp. 639–667.
doi:10.1007/s11948-010-9201-y.

[45] Zanoni Dias, Anderson Rocha, and Siome Goldenstein. “First Steps Toward Image Phylogeny”. In: *IEEE Workshop on Information Forensics and Security, WIFS 2010* (Seattle, WA, Dec. 12–15, 2010). 2010.
doi:10.1109/WIFS.2010.5711452.

[46] A. Emir Dirik and Nasir Memon. “Image Tamper Detection Based on Demosaicing Artifacts”. In: *IEEE International Conference on Image Processing, ICIP 2009* (Cairo, Egypt, Nov. 7–10, 2009). 2009, pp. 1497–1500.
doi:10.1109/ICIP.2009.5414611.

[47] A. Emir Dirik, Sevinç Bayram, Husrev T. Sencar, and Nasir Memon. “New Features to Identify Computer Generated Images”. In: *IEEE International Conference on Image Processing, ICIP 2007* (San Antonio, TX, Sept. 16–19, 2007). Vol. 4. 2007, pp. IV–433–436.
doi:10.1109/ICIP.2007.4380047.

[48] A. Emir Dirik, Husrev T. Sencar, and Nasir Memon. “Digital Single Lens Reflex Camera Identification From Traces of Sensor Dust”. In: *IEEE Transactions on Information Forensics and Security* 3.3 (Sept. 2009), pp. 539–552.
doi:10.1109/TIFS.2008.926987.

[49] Brandon Dybala, Brian Jennings, and David Letscher. “Detecting Filtered Cloning in Digital Images”. In: *MM&Sec’07, Proceedings of the Multimedia and Security Workshop 2007* (Dallas, TX, Sept. 20–21, 2007). New York, NY: ACM Press, 2007, pp. 43–50.
doi:10.1145/1288869.1288877.

[50] Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa. “PhotoSketch: A Sketch Based Image Query and Compositing System”. In: *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference (SIGGRAPH) 2009* (New Orleans, LA, Aug. 3–7, 2009). New York, NY: ACM Press, 2009.
doi:10.1145/1597990.1598050.

[51] Ahmet M. Eskicioglu and Paul S. Fisher. “Image Quality Measures and Their Performance”. In: *IEEE Transactions on Communications* 43.12 (Dec. 1995), pp. 2959–2965.
doi:10.1109/26.477498.

[52] Zhigang Fan and Ricardo L. de Queiroz. “Identification of Bitmap Compression History: JPEG Detection and Quantizer Estimation”. In: *IEEE Transactions on Image Processing* 12.2 (Feb. 2003), pp. 230–235.
doi:10.1109/TIP.2002.807361.

[53] Yanmei Fang, A. Emir Dirik, Xiaoxi Sun, and Nasir Memon. “Source Class Identification for DSLR and Compact Cameras”. In: *IEEE International Workshop on Multimedia Signal Processing 2009, MMSP’09* (Rio de Janeiro, Brazil, Oct. 5–7, 2009). 2009.
doi:10.1109/MMSP.2009.5293342.

[54] Hany Farid. *Detecting Digital Forgeries Using Bispectral Analysis*. Technical Report AIM-1657. AI Lab, Massachusetts Institute of Technology, 1999.
ftp://publications.ai.mit.edu/ai-publications/pdf/AIM-1657.pdf.

[55] Hany Farid. *Digital Image Ballistics from JPEG Quantization: A Followup Study*. Tech. rep. TR2008-638. Hanover, NH: Department of Computer Science, Dartmouth College, 2008.
http://www.cs.dartmouth.edu/reports/TR2008-638.pdf.

[56] Hany Farid. “Exposing Digital Forgeries From JPEG Ghosts”. In: *IEEE Transactions on Information Forensics and Security* 4.1 (Mar. 2009), pp. 154–160.
doi:10.1109/TIFS.2008.2012215.

[57] Hany Farid. “Image Forgery Detection”. In: *IEEE Signal Processing Magazine* 26.2 (Mar. 2009): *Digital Forensics*, pp. 16–25.
doi:10.1109/MSP.2008.931079.

[58] Hany Farid. *Photo Tampering Throughout History*. 2011.
http://www.cs.dartmouth.edu/farid/research/digitaltampering.

[59] Hany Farid and Mary J. Bravo. “Image Forensic Analyses that Elude the Human Visual System”. In: *Media Forensics and Security II* (San Jose, CA, Jan. 18–20, 2010). Ed. by Nasir D. Memon, Jana Dittmann, Adnan M. Alattar, and Edward J. Delp. Vol. 7541. Proceedings of SPIE. Bellingham, WA: SPIE, 2010, 754106.
doi:10.1117/12.837788.

[60] Hany Farid and Siwei Lyu. “Higher-Order Wavelet Statistics and their Application to Digital Forensics”. In: *IEEE Workshop on Statistical Analysis in Computer Vision (in conjunction with CVPR)* (Madison, WI, June 22, 2003). 2003.
doi:10.1109/CVPRW.2003.10093.

[61] Xiaoying Feng, Gwenaël Doërr, and Ingemar J. Cox. “An Energy-Based Method for the Forensic Detection of Re-sampled Images”. In: *IEEE International Conference on Multimedia and EXPO, ICME 2011* (Barcelona, Spain, July 11–15, 2011). 2011. doi:10.1109/ICME.2011.6011984.

[62] David J. Field. “Relations Between the Statistics of Natural Images and the Response Properties of Cortical Cells”. In: *Journal of the Optical Society of America A* 4.12 (1987), pp. 2379–2394. doi:10.1364/JOSAA.4.002379.

[63] Tomáš Filler, Jessica Fridrich, and Miroslav Goljan. “Using Sensor Pattern Noise for Camera Model Identification”. In: *2008 IEEE International Conference on Image Processing, ICIP 2008* (San Diego, CA, Oct. 12–15, 2008). 2008, pp. 1296–1299. doi:10.1109/ICIP.2008.4712000.

[64] Katrin Y. Franke and Sargur N. Srihari. “Computational Forensics: An Overview”. In: *Computational Forensics, Second International Workshop, IWCF 2008* (Washington, DC, Aug. 7–8, 2008). Ed. by Sargur N. Srihari and Katrin Y. Franke. Lecture Notes in Computer Science 5158. Berlin, Heidelberg: Springer Verlag, 2008, pp. 1–10. doi:10.1007/978-3-540-85303-9_1.

[65] Jessica Fridrich. “Digital Image Forensics”. In: *IEEE Signal Processing Magazine* 26.2 (Mar. 2009): *Digital Forensics*, pp. 26–37. doi:10.1109/MSP.2008.931078.

[66] Jessica Fridrich. *Steganography in Digital Media. Principles, Algorithms, and Applications*. New York: Cambridge University Press, 2010.

[67] Jessica Fridrich and Miroslav Goljan. “Determining Approximate Age of Digital Images Using Sensor Defects”. In: *Media Watermarking, Security, and Forensics III* (San Francisco, CA, Jan. 24–26, 2011). Ed. by Nasir D. Memon, Jana Dittmann, Adnan M. Alattar, and Edward J. Delp. Vol. 7880. Proceedings of SPIE. Bellingham, WA: SPIE Press, 2011, 788006. doi:10.1117/12.872198.

[68] Jessica Fridrich, Miroslav Goljan, and Rui Du. “Steganalysis Based on JPEG Compatibility”. In: *Multimedia Systems and Applications IV* (Denver, CO, Aug. 21–22, 2001). Ed. by Andrew G. Tescher, Bhaskaran Vasudev, and V. Michael Bove. Vol. 4518. Proceedings of SPIE. Bellingham, WA: SPIE Press, 2001, pp. 275–280. doi:10.1117/12.448213.

[69] Jessica Fridrich, David Soukal, and Jan Lukáš. “Detection of Copy-Move Forgery in Digital Images”. In: *Digital Forensic Research Workshop* (Cleveland, OH, Aug. 6–8, 2003). 2003. http://www.ws.binghamton.edu/fridrich/Research/copymove.pdf.

[70] Jessica Fridrich, Jan Kodovský, Vojtěch Holub, and Miroslav Goljan. “Breaking HUGO—The Process Discovery”. In: *Information Hiding, 13th International Conference, IH 2011. Revised Selected Papers* (Prague, Czech Republic, May 18–20, 2011). Ed. by Tomáš Filler, Tomáš Pevný, Scott Craver, and Andrew Ker. Vol. 6958. Berlin, Heidelberg: Springer Verlag, 2011, pp. 85–101. doi:10.1007/978-3-642-24178-9_7.

[71] Gary L. Friedman. “The Trustworthy Digital Camera: Restoring Credibility to the Photographic Image”. In: *IEEE Transactions on Consumer Electronics* 39.4 (Nov. 1993), pp. 905–910.
doi:10.1109/30.267415.

[72] Dongdong Fu, Yun Q. Shi, and Wei Su. “A Generalized Benford’s Law for JPEG Coefficients and its Applications in Image Forensics”. In: *Security and Watermarking of Multimedia Content IX* (San Jose, CA, Jan. 29–Feb. 1, 2007). Ed. by Edward J. Delp and Ping Wah Wong. Vol. 6505. Proceedings of SPIE. Bellingham, WA: SPIE, 2007, 65051L.
doi:10.1117/12.704723.

[73] Teddy Furon and Patrick Bas. “Broken Arrows”. In: *EURASIP Journal on Information Security* 2008.597040 (2008).
doi:10.1155/2008/597040.

[74] Andrew C. Gallagher. “Detection of Linear and Cubic Interpolation in JPEG Compressed Images”. In: *Second Canadian Conference on Computer and Robot Vision, CRV 2005* (Victoria, BC, May 9–11, 2005). 2005, pp. 65–72.
doi:10.1109/CRV.2005.33.

[75] Andrew C. Gallagher and Tsu-Han Chen. “Image Authentication by Detecting Traces of Demosaicing”. In: *IEEE Workitorial on Vision of the Unseen (in conjunction with CVPR)* (Anchorage, AK, June 23, 2008). 2008.
doi:10.1109/CVPRW.2008.4562984.

[76] Xinting Gao, Tian-Tsong Ng, Bo Qiu, and Shih-Fu Chang. “Single-View Recaptured Image Detection Based on Physics-Based Features”. In: *IEEE International Conference on Multimedia and EXPO, ICME 2010* (Singapore, July 19–23, 2010). 2010, pp. 1469–1474.
doi:10.1109/ICME.2010.5583280.

[77] Matthew D. Gaubatz and Steven J. Simske. “Printer-Scanner Identification via Analysis of Structured Security Deterrents”. In: *IEEE International Workshop on Information Forensics and Security, WIFS 2009* (London, UK, Dec. 6–9, 2009). 2009, pp. 151–155.
doi:10.1109/WIFS.2009.5386463.

[78] Matthew D. Gaubatz and Steven J. Simske. “Towards a Feature Set for Robust Printing-Imaging Cycle Device Identification Using Structured Printed Markings”. In: *IEEE Workshop on Information Forensics and Security, WIFS 2010* (Seattle, WA, Dec. 12–15, 2010). 2010.
doi:10.1109/WIFS.2010.5711463.

[79] Zeno J. Geradts, Jurrien Bijhold, Martijn Kieft, Kenji Kurosawa, Kenro Kuroki, and Naoki Saitoh. “Methods for Identification of Images Acquired with Digital Cameras”. In: *Enabling Technologies for Law Enforcement and Security* (Boston, MA, Nov. 5, 2000). Ed. by Simon K. Bramble, Edward M. Carapezza, and Lenny I. Rudin. Vol. 4232. Proceedings of SPIE. Bellingham, WA: SPIE, 2001, pp. 505–512.
doi:10.1117/12.417569.

[80] Bernd Girod. “What’s Wrong with Mean-Squared Error?” In: *Digital Images and Human Vision*. Ed. by Andrew B. Watson. MIT Press, 1993, pp. 207–220.

[81] Thomas Gloe. “Demystifying Histograms of Multi-Quantized DCT Coefficients”. In: *IEEE International Conference on Multimedia and EXPO, ICME 2011* (Barcelona, Spain, July 11–15, 2011). 2011.

[82] Thomas Gloe and Rainer Böhme. “The Dresden Image Database for Benchmarking Digital Image Forensics”. In: *Journal of Digital Forensic Practice* 3.2–4 (2010), pp. 150–159.
doi:10.1080/15567281.2010.531500.

[83] Thomas Gloe, Matthias Kirchner, Antje Winkler, and Rainer Böhme. “Can we Trust Digital Image Forensics?” In: *MULTIMEDIA ’07, Proceedings of the 15th International Conference on Multimedia* (Augsburg, Germany, Sept. 24–29, 2007). New York, NY: ACM Press, 2007, pp. 78–86.
doi:10.1145/1291233.1291252.

[84] Thomas Gloe, Elke Franz, and Antje Winkler. “Forensics for Flatbed Scanners”. In: *Security and Watermarking of Multimedia Content IX* (San Jose, CA, Jan. 29–Feb. 1, 2007). Ed. by Edward J. Delp and Ping Wah Wong. Vol. 6505. Proceedings of SPIE. Bellingham, WA: SPIE, 2007, 65051I.
doi:10.1117/12.704165.

[85] Thomas Gloe, Karsten Borowka, and Antje Winkler. “Feature-Based Camera Model Identification Works in Practice: Results of a Comprehensive Evaluation Study”. In: *Information Hiding, 11th International Workshop, IH 2009. Revised Selected Papers* (Darmstadt, Germany, June 8–10, 2009). Ed. by Stefan Katzenbeisser and Ahmad-Reza Sadeghi. Vol. 5806. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 262–276.
doi:10.1007/978-3-642-04431-1_19.

[86] Thomas Gloe, Karsten Borowka, and Antje Winkler. “Efficient Estimation and Large-Scale Evaluation of Lateral Chromatic Aberration for Digital Image Forensics”. In: *Media Forensics and Security II* (San Jose, CA, Jan. 18–20, 2010). Ed. by Nasir D. Memon, Jana Dittmann, Adnan M. Alattar, and Edward J. Delp. Vol. 7541. Proceedings of SPIE. Bellingham, WA: SPIE, 2010, 754107.
doi:10.1117/12.839034.

[87] Miroslav Goljan, Mo Chen, and Jessica Fridrich. “Identifying Common Source Digital Camera from Image Pairs”. In: *IEEE International Conference on Image Processing, ICIP 2007* (San Antonio, TX, Sept. 16–19, 2007). Vol. 6. 2007, pp. VI–125–128.
doi:10.1109/ICIP.2007.4379537.

[88] Miroslav Goljan, Jessica Fridrich, and Jan Lukáš. “Camera Identification from Printed Images”. In: *Security, Forensics, Steganography, and Watermarking of Multimedia Contents X* (San Jose, CA, Jan. 28–30, 2008). Ed. by Edward J. Delp, Ping Wah Wong, Jana Dittmann, and Nasir D. Memon. Vol. 6819. Proceedings of SPIE. Bellingham, WA: SPIE, 2008, 68190I.
doi:10.1117/12.766824.

[89] Miroslav Goljan, Jessica Fridrich, and Tomáš Filler. “Large Scale Test of Sensor Fingerprint Camera Identification”. In: *Media Forensics and Security* (San Jose, CA, Jan. 19–21, 2009). Ed. by Edward J. Delp, Jana Dittmann, Nasir D. Memon, and Ping Wah Wong. Vol. 7254. Proceedings of SPIE. Bellingham, WA: SPIE, 2009, 72540I.
doi:10.1117/12.805701.

[90] Miroslav Goljan, Jessica Fridrich, and Tomáš Filler. “Managing a Large Database of Camera Fingerprints”. In: *Media Forensics and Security II* (San Jose, CA, Jan. 18–20, 2010). Ed. by Nasir D. Memon, Jana Dittmann, Adnan M. Alattar, and Edward J. Delp. Vol. 7541. Proceedings of SPIE. Bellingham, WA: SPIE, 2010, 754108.
doi:10.1117/12.838378.

[91] Miroslav Goljan, Jessica Fridrich, and Mo Chen. “Sensor Noise Camera Identification: Countering Counter-Forensics”. In: *Media Forensics and Security II* (San Jose, CA, Jan. 18–20, 2010). Ed. by Nasir D. Memon, Jana Dittmann, Adnan M. Alattar, and Edward J. Delp. Vol. 7541. Proceedings of SPIE. Bellingham, WA: SPIE, 2010, 75410S.
doi:10.1117/12.839055.

[92] Miroslav Goljan, Jessica Fridrich, and Mo Chen. “Defending Against Fingerprint-Copy Attack in Sensor-Based Camera Identification”. In: *IEEE Transactions on Information Forensics and Security* 6.1 (Mar. 2011), pp. 227–236.
doi:10.1109/TIFS.2010.2099220.

[93] Hongmei Gou, Ashwin Swaminathan, and Min Wu. “Intrinsic Sensor Noise Features for Forensic Analysis on Scanners and Scanned Images”. In: *IEEE Transactions on Information Forensics and Security* 4.3 (Sept. 2009), pp. 476–491.
doi:10.1109/TIFS.2009.2026458.

[94] Catalin Grigoras. “Applications of ENF Criterion in Forensic Audio, Video, Computer and Telecommunication Analysis”. In: *Forensic Science International* 167.2–3 (Apr. 2007), pp. 136–145.
doi:10.1016/j.forsciint.2006.06.033.

[95] Ryan Harris. “Arriving at an Anti-Forensics Consensus: Examining how to Define and Control the Anti-Forensics Problem”. In: *Digital Investigation* 3.S1 (Sept. 2006): *The Proceedings of the 6th Annual Digital Forensic Research Workshop (DFRWS ’06)*, pp. 44–49.
doi:10.1016/j.diin.2006.06.005.

[96] Glenn E. Healey and Raghava Kondepudy. “Radiometric CCD Camera Calibration and Noise Estimation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16.3 (Mar. 1994), pp. 267–276.
doi:10.1109/34.276126.

[97] Volker Heerich. “Die Identifikation von Faxgeräten [Identfication of Fax Machines]”. In German. In: *Kriminalistik* 52.3 (Mar. 1998), pp. 214–217.

[98] Martin E. Hellman and Josef Raviv. “Probability of Error, Equivocation, and the Chernoff Bound”. In: *IEEE Transactions on Information Theory* 16.4 (July 1970), pp. 368–372.
doi:10.1109/TIT.1970.1054466.

[99] John Ho, Oscar C. Au, and Jiantao Zhou. “Inter-Channel Demosaicking Traces for Digital Image Forensics”. In: *IEEE International Conference on Multimedia and EXPO, ICME 2010* (Singapore, July 19–23, 2010). 2010, pp. 1475–1480. doi:10.1109/ICME.2010.5582951.

[100] Gerald C. Holst and Terrence S. Lomheim. *CMOS/CCD Sensors and Camera Systems*. Bellingham, WA: SPIE Press, 2007.

[101] Chih-Chung Hsu, Tzu-Yi Hung, Chia-Wen Lin, and Chiou-Ting Hsu. “Video Forgery Detection Using Correlation of Noise Residue”. In: *IEEE 10th Workshop on Multimedia Signal Processing (MMSP)* (Cairns, Qld, Oct. 8–10, 2008). 2008, pp. 170–174. doi:10.1109/MMSP.2008.4665069.

[102] Yu-Feng Hsu and Shih-Fu Chang. “Detecting Image Splicing using Geometry Invariants and Camera Characteristics Consistency”. In: *IEEE International Conference on Multimedia and EXPO, ICME 2006* (Toronto, ON, July 9–12, 2006). 2006, pp. 549–552.

[103] Yu-Feng Hsu and Shih-Fu Chang. “Image Splicing Detection using Camera Response Function Consistency and Automatic Segmentation”. In: *IEEE International Conference on Multimedia and EXPO, ICME 2007* (Beijing, China, July 2–5, 2007). 2007, pp. 28–31. doi:10.1109/ICME.2007.4284578.

[104] Fangjun Huang, Jiwu Huang, and Yun Qing Shi. “Detecting Double JPEG Compression With the Same Quantization Matrix”. In: *IEEE Transactions on Information Forensics and Security* 5.4 (Dec. 2010), pp. 848–856. doi:10.1109/TIFS.2010.2072921.

[105] Hailing Huang, Weiqiang Guo, and Yu Zhang. “Detection of Copy-Move Forgery in Digital Images Using SIFT Algorithm”. In: *Pacific-Asia Workshop on Computational Intelligence and Industrial Application* (Wuhan, China, Dec. 19–20, 2008). Vol. 2. 2008, pp. 272–276. doi:10.1109/PACIIA.2008.240.

[106] Yizhen Huang. “Can Digital Image Forgery Detection be Unevadable? A Case Study: Color Filter Array Interpolation Statistical Feature Recovery”. In: *Visual Communications and Image Processing* (Beijing, China, July 12–15, 2005). Ed. by Shipeng Li, Fernando Pereira, Heung-Yeung Shum, and Andrew G. Tescher. Vol. 5960. Proceedings of SPIE. Bellingham, WA: SPIE, 2005, 59602W. doi:10.1117/12.632564.

[107] Aapo Hyvärinen, Jarmo Hurri, and Patrick O. Hoyer. *Natural Image Statistics. A Probabilistic Approach to Early Computational Vision.* London: Springer-Verlag, 2009.

[108] ISO/IEC 10918-1. *Information Technology. Digital Compression and Coding of Continuous-tone Still Images. International Standard*. 1994.

[109] ISO/IEC 15444-1. *Information Technology. JPEG 2000 Image Coding System: Core Coding System. International Standard*. 2004.

[110] Japan Electronics and Information Technology Industries Association. *Exchangeable Image File Format for Digital Still Cameras: Exif Version 2.2*. Apr. 2002. http://www.exif.org/Exif2-2.PDF.

[111] Micah K. Johnson and Hany Farid. “Exposing Digital Forgeries through Chromatic Aberration”. In: *MM&Sec’06, Proceedings of the Multimedia and Security Workshop 2006* (Geneva, Switzerland, Sept. 26–27, 2006). New York, NY: ACM Press, 2006, pp. 48–55. doi:10.1145/1161366.1161376.

[112] Micah K. Johnson and Hany Farid. “Exposing Digital Forgeries in Complex Lighting Environments”. In: *IEEE Transactions on Information Forensics and Security* 2.3 (Sept. 2007), pp. 450–461. doi:10.1109/TIFS.2007.903848.

[113] Thomas Kailath. “The Divergence and Bhattacharyya Distance Measures in Signal Selection”. In: *IEEE Transactions on Communication Technology* 15.1 (Feb. 1967), pp. 52–60. doi:10.1109/TCOM.1967.1089532.

[114] Ton Kalker. “Considerations on Watermarking Security”. In: *IEEE Fourth Workshop on Multimedia Signal Processing* (Cannes, France, Oct. 3–5, 2001). 2001, pp. 201–206. doi:10.1109/MMSP.2001.962734.

[115] Steven M. Kay. *Fundamentals of Statistical Signal Processing. Detection Theory*. Upper Saddle River, NJ: Prentice Hall, 1998.

[116] Eric Kee and Hany Farid. “Printer Profiling for Forensics and Ballistics”. In: *MM&Sec’08, Proceedings of the Multimedia and Security Workshop 2008* (Oxford, UK, Sept. 22–23, 2008). New York, NY: ACM Press, 2008, pp. 3–10. doi:10.1145/1411328.1411332.

[117] Eric Kee and Hany Farid. “Digital Image Authentication from Thumbnails”. In: *Media Forensics and Security II* (San Jose, CA, Jan. 18–20, 2010). Ed. by Nasir D. Memon, Jana Dittmann, Adnan M. Alattar, and Edward J. Delp. Vol. 7541. Proceedings of SPIE. Bellingham, WA: SPIE, 2010, 75410E. doi:10.1117/12.838834.

[118] Eric Kee, Micah K. Johnson, and Hany Farid. “Digital Image Authentication from JPEG Headers”. In: *IEEE Transactions on Information Forensics and Security* 6.3 (Sept. 2011), pp. 1066–1075. doi:10.1109/TIFS.2011.2128309.

[119] John Kelsey, Bruce Schneier, and Chris Hall. “An Authenticated Camera”. In: *12th Annual Computer Security Applications Conference (ACSAC’96)* (San Diego, CA, Dec. 9–13, 1996). 1996, pp. 24–30. doi:10.1109/CSAC.1996.569666.

[120] Auguste Kerckhoffs. “La cryptographie militaire”. In: *Journal des sciences militaires* IX (1883), pp. 5–38, 161–191. http://www.petitcolas.net/fabien/kerckhoffs/crypto_militaire_1.pdf.

[121] Nitin Khanna, Aravind K. Mikkilineni, Anthony F. Martone, Gazi N. Ali, George T.-C. Chiu, Jan P. Allebach, and Edward J. Delp. “A Survey of Forensic Characterization Methods for Physical Devices”. In: *Digital Investigation* 3.S1 (Sept. 2006): *The Proceedings of the 6th Annual Digital Forensic Research Workshop (DFRWS ’06)*, pp. 17–28. doi:10.1016/j.diin.2006.06.014.

[122] Nitin Khanna, Aravind K. Mikkikineni, George T.-C. Chiu, Jan P. Allebach, and Edward J. Delp. “Scanner Identification Using Sensor Pattern Noise”. In: *Security and Watermarking of Multimedia Content IX* (San Jose, CA, Jan. 29–Feb. 1, 2007). Ed. by Edward J. Delp and Ping Wah Wong. Vol. 6505. Proceedings of SPIE. Bellingham, WA: SPIE, 2007, 65051K.
doi:10.1117/12.705837.

[123] Nitin Khanna, George T.-C. Chiu, Jan P. Allebach, and Edward J. Delp. “Forensic Techniques for Classifying Scanner, Computer Generated and Digital Camera Images”. In: *2008 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2008* (Las Vegas, NV, Mar. 31–Apr. 4, 2008). 2008, pp. 1653–1656.
doi:10.1109/ICASSP.2008.4517944.

[124] Nitin Khanna, Aravind K. Mikkilineni, and Edward J. Delp. “Scanner Identification Using Feature-Based Processing and Analysis”. In: *IEEE Transactions on Information Forensics and Security* 4.1 (Mar. 2009), pp. 123–139.
doi:10.1109/TIFS.2008.2009604.

[125] Mehdi Kharrazi, Husrev T. Sencar, and Nasir Memon. “Blind Source Camera Identification”. In: *IEEE International Conference on Image Processing, ICIP 2004* (Singapore, Oct. 24–27, 2004). Vol. 1. 2004, pp. 709–712.
doi:10.1109/ICIP.2004.1418853.

[126] Matthias Kirchner. “Fast and Reliable Resampling Detection by Spectral Analysis of Fixed Linear Predictor Residue”. In: *MM&Sec’08, Proceedings of the Multimedia and Security Workshop 2008* (Oxford, UK, Sept. 22–23, 2008). New York, NY: ACM Press, 2008, pp. 11–20.
doi:10.1145/1411328.1411333.

[127] Matthias Kirchner. “Efficient Estimation of CFA Pattern Configuration in Digital Camera Images”. In: *Media Forensics and Security II* (San Jose, CA, Jan. 18–20, 2010). Ed. by Nasir D. Memon, Jana Dittmann, Adnan M. Alattar, and Edward J. Delp. Vol. 7541. Proceedings of SPIE. Bellingham, WA: SPIE, 2010, 754111.
doi:10.1117/12.839102.

[128] Matthias Kirchner and Rainer Böhme. “Tamper Hiding: Defeating Image Forensics”. In: *Information Hiding, 9th International Workshop, IH 2007. Revised Selected Papers* (Saint Malo, France, June 11–13, 2007). Ed. by Teddy Furon, François Cayre, Gwenaël Doërr, and Patrick Bas. Vol. 4567. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 326–341.
doi:10.1007/978-3-540-77370-2_22.

[129] Matthias Kirchner and Rainer Böhme. “Hiding Traces of Resampling in Digital Images”. In: *IEEE Transactions on Information Forensics and Security* 3.4 (Dec. 2008), pp. 582–592.
doi:10.1109/TIFS.2008.2008214.

[130] Matthias Kirchner and Rainer Böhme. “Synthesis of Color Filter Array Pattern in Digital Images”. In: *Media Forensics and Security* (San Jose, CA, Jan. 19–21, 2009). Ed. by Edward J. Delp, Jana Dittmann, Nasir D. Memon, and Ping Wah Wong. Vol. 7254. Proceedings of SPIE. Bellingham, WA: SPIE, 2009, 72540K.

doi:10.1117/12.805988.

[131] Matthias Kirchner and Jessica Fridrich. “On Detection of Median Filtering in Digital Images”. In: *Media Forensics and Security II* (San Jose, CA, Jan. 18–20, 2010). Ed. by Nasir D. Memon, Jana Dittmann, Adnan M. Alattar, and Edward J. Delp. Vol. 7541. Proceedings of SPIE. Bellingham, WA: SPIE, 2010, 754110.

[132] Matthias Kirchner and Thomas Gloe. “On Resampling Detection in Re-compressed Images”. In: *2009 First IEEE International Workshop on Information Forensics and Security, WIFS 2009* (London, UK, Dec. 6–9, 2009). 2009, pp. 21–25.
doi:10.1109/WIFS.2009.5386489.

[133] Simon Knight, Simon Moschou, and Matthew Sorell. “Analysis of Sensor Photo Response Non-Uniformity in RAW Images”. In: *Forensics in Telecommunications, Information and Multimedia, Second International Conference, e-Forensics 2009. Revised Selected Papers* (Adelaide, SA, Jan. 19–21, 2009). Ed. by Matthew Sorell. Vol. 8. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 130–141.
doi:10.1007/978-3-642-02312-5_15.

[134] Michael Knopp. “Digitalfotos als Beweismittel [Digital Images as Pieces of Evidence]”. In German. In: *Zeitschrift für Rechtspolitik* 41.5 (2008), pp. 156–158.

[135] Jan Kodovský and Jessica Fridrich. “Steganalysis in High Dimensions: Fusing Classifiers Built on Random Subspaces”. In: *Media Watermarking, Security, and Forensics III* (San Francisco, CA, Jan. 24–26, 2011). Ed. by Nasir D. Memon, Jana Dittmann, Adnan M. Alattar, and Edward J. Delp. Vol. 7880. Proceedings of SPIE. Bellingham, WA: SPIE, 2011, 78800L.
doi:10.1117/12.872279.

[136] Solomon Kullback and Richard Leibler. “On Information and Sufficiency”. In: *Annals of Mathematical Statistics* 22.1 (Mar. 1951), pp. 79–86.
doi:10.1214/aoms/1177729694.

[137] Kenji Kurosawa, Kenro Kuroki, and Naoki Saitoh. “CCD Fingerprint Method – Identification of a Video Camera from Videotaped Images”. In: *IEEE International Conference on Image Processing, ICIP 1999* (Kobe, Japan, Oct. 24–28, 1999). Vol. 3. 1999, pp. 537–540.
doi:10.1109/ICIP.1999.817172.

[138] Kenji Kurosawa, Kenro Kuroki, and Norimitsu Akiba. “Individual Camera Identification Using Correlation of Fixed Pattern Noise in Image Sensors”. In: *Journal of Forensic Sciences* 54.3 (May 2009), pp. 639–641.
doi:10.1111/j.1556-4029.2009.01017.x.

[139] ShiYue Lai and Rainer Böhme. “Countering Counter-Forensics: The Case of JPEG Compression”. In: *Information Hiding, 13th International Conference, IH 2011. Revised Selected Papers* (Prague, Czech Republic, May 18–20, 2011). Ed. by Tomáš Filler, Tomáš Pevný, Scott Craver, and Andrew Ker. Vol. 6958. Berlin, Heidelberg: Springer Verlag, 2011, pp. 285–298.
doi:10.1007/978-3-642-24178-9_20.

[140] Jean-François Lalonde, Derek Hoeim, Alexei A. Efros, Carsten Rother, John Winn, and Antonio Criminisi. “Photo Clip Art”. In: *ACM Transactions on Graphics* 26.3 (Aug. 2007): *Proceedings of ACM SIGGRAPH 2007*.
doi:10.1145/1276377.1276381.

[141] Ji-Won Lee, Min-Jeong Lee, Tae-Woo Oh, Seung-Jin Ryu, and Heung-Kyu Lee. “Screenshot Identification Using Combing Artifact from Interlaced Video”. In: *MM&Sec'10, Proceedings of the 2010 ACM SIGMM Multimedia and Security Workshop* (Rome, Italy, Sept. 9–10, 2010). New York: ACM Press, 2010, pp. 49–54.
doi:10.1145/1854229.1854240.

[142] Chang-Tsun Li, Chih-Yuan Chang, and Yue Li. “On the Repudiability of Device Identification and Image Integrity Verification Using Sensor Pattern Noise”. In: *Information Security and Digital Forensics, First International Conference, ISDF 2009. Revised Selected Papers* (London, UK, Sept. 7–9, 2009). Ed. by Dasun Weerasinghe. Vol. 41. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Berlin, Heidelberg: Springer, 2010, pp. 19–25.
doi:10.1007/978-3-642-11530-1_3.

[143] Weihai Li, Yuan Yuan, and Nenghai Yu. “Passive Detection of Doctored JPEG Image via Block Artifact Grid Extraction”. In: *Signal Processing* 89.9 (Sept. 2009), pp. 1821–1829.
doi:10.1016/j.sigpro.2009.03.025.

[144] Yue Li and Chang-Tsun Li. “Decomposed Photo Response Non-Uniformity for Digital Forensic Analysis”. In: *Forensics in Telecommunications, Information and Multimedia, Second International Conference, e-Forensics 2009. Revised Selected Papers* (Adelaide, SA, Jan. 19–21, 2009). Ed. by Matthew Sorell. Vol. 8. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 166–172.
doi:10.1007/978-3-642-02312-5_19.

[145] Jianhua Lin. “Divergence Measures Based on the Shannon Entropy”. In: *IEEE Transactions on Information Theory* 37.1 (Jan. 1991), pp. 145–151.
doi:10.1109/18.61115.

[146] W. Sabrina Lin, Steven Tjoa, H. Vicky Zhao, and K. J. Ray Liu. “Digital Image Source Coder Forensics via Intrinsic Fingerprints”. In: *IEEE Transactions on Information Forensics and Security* 4.3 (Sept. 2009), pp. 460–475.
doi:10.1109/TIFS.2009.2024715.

[147] Zhouchen Lin, Rongrong Wang, Xiaoou Tang, and Heung-Yeung Shum. “Detecting Doctored Images using Camera Response Normality and Consistency”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005* (San Diego, CA, June 20–26, 2005). Vol. 1. 2005, pp. 1087–1092.
doi:10.1109/CVPR.2005.125.

[148] Zhouchen Lin, Junfeng He, Xiaoou Tang, and Chi-Keung Tang. “Fast, Automatic and Fine-grained Tampered JPEG Image Detection via DCT Coefficient Analysis”. In: *Pattern Recognition* 42.11 (Nov. 2009), pp. 2492–2501.
doi:10.1016/j.patcog.2009.03.019.

[149] Qiguang Liu, Xiaochun Cao, Chao Deng, and Xiaojie Guo. “Identifying Image Composites through Shadow Matte Consistency”. In: *IEEE Transactions on Information Forensics and Security* 6.3 (Sept. 2011), pp. 1111–1122.
doi:10.1109/TIFS.2011.2139209.

[150] Gareth A. Lloyd and Steven J. Sasson. *Electronic Still Camera*. US Patent, 4 131 919. 1978.

[151] Jan Lukáš. “Digital Image Authentication Using Image Filtering Techniques”. In: *Algoritmy 2000. Proceedings of Contributed Papers and Posters* (Podbanske, Slovakia, Sept. 10–15, 2000). 2000, pp. 236–244.
http://www.emis.de/journals/AMUC/_contributed/algo2000/lukas.pdf.

[152] Jan Lukáš and Jessica Fridrich. “Estimation of Primary Quantization Matrix in Double Compressed JPEG Images”. In: *Digital Forensic Research Workshop* (Cleveland, OH, Aug. 6–8, 2003). 2003.
http://www.ws.binghamton.edu/fridrich/Research/Doublecompression.pdf.

[153] Jan Lukáš, Jessica Fridrich, and Miroslav Goljan. “Determining Digital Image Origin Using Sensor Imperfections”. In: *Image and Video Communications and Processing* (San Jose, CA, Jan. 18–20, 2005). Ed. by Amir Said and John G. Apostolopoulos. Vol. 5685. Proceedings of SPIE. Bellingham, WA: SPIE, 2005, pp. 249–260.
doi:10.1117/12.587105.

[154] Jan Lukáš, Jessica Fridrich, and Miroslav Goljan. “Detecting Digital Image Forgeries Using Sensor Pattern Noise”. In: *Security and Watermarking of Multimedia Content VIII* (San Jose, CA, Jan. 16–19, 2006). Ed. by Edward J. Delp and Ping Wah Wong. Vol. 6072. Proceedings of SPIE. Bellingham, WA: SPIE, 2006, 60720Y.
doi:10.1117/12.640109.

[155] Jan Lukáš, Jessica Fridrich, and Miroslav Goljan. “Digital Camera Identification from Sensor Pattern Noise”. In: *IEEE Transactions on Information Forensics and Security* 1.2 (June 2006), pp. 205–214.
doi:10.1109/TIFS.2006.873602.

[156] Weiqi Luo, Yuangen Wang, and Jiwu Huang. “Detection of Quantization Artifacts and Its Applications to Transform Encoder Identification”. In: *IEEE Transactions on Information Forensics and Security* 5.4 (Dec. 2010), pp. 810–815.
doi:10.1109/TIFS.2010.2074195.

[157] Weiqi Luo, Jiwu Huang, and Guoping Qiu. “JPEG Error Analysis and Its Applications to Digital Image Forensics”. In: *IEEE Transactions on Information Forensics and Security* 5.3 (Sept. 2010), pp. 480–491.
doi:10.1109/TIFS.2010.2051426.

[158] Siwei Lyu. “Estimating Vignetting Function from a Single Image for Image Authentication”. In: *MM&Sec’10, Proceedings of the 2010 ACM SIGMM Multimedia & Security Workshop* (Rome, Italy, Sept. 9–10, 2010). New York: ACM Press, 2010, pp. 3–12.
doi:10.1145/1854229.1854233.

[159] Siwei Lyu and Hany Farid. “How Realistic is Photorealistic?” In: *IEEE Transactions on Signal Processing* 53.2 (Feb. 2005), pp. 845–850.

doi:10.1109/TSP.2004.839896.

[160] Babak Mahdian and Stanislav Saic. “Detection of Copy-Move Forgery Using a Method Based on Blur Moment Invariants”. In: *Forensic Science International* 171.2–3 (Sept. 2007), pp. 180–189.
doi:10.1016/j.forsciint.2006.11.002.

[161] Babak Mahdian and Stanislav Saic. “Blind Methods for Detecting Image Fakery”. In: *IEEE Aerospace and Electronic Systems Magazine* 25.4 (Apr. 2010), pp. 18–24.
doi:10.1109/MAES.2010.5467652.

[162] Junwen Mao, Orhan Bulan, Gaurav Sharma, and Suprakash Datta. “Device Temporal Forensics: An Information Theoretic Approach”. In: *IEEE International Conference on Image Processing, ICIP 2009* (Cairo, Egypt, Nov. 7–10, 2009). 2009, pp. 1501–1504.
doi:10.1109/ICIP.2009.5414612.

[163] Emma Marris. “Should Journals Police Scientific Fraud?” In: *Nature* 439.2 (Feb. 2006), pp. 520–521.
doi:10.1038/439520a.

[164] Christine McKay, Ashwin Swaminathan, Hongmei Gou, and Min Wu. “Image Acquisition Forensics: Forensic Analysis to Identify Imaging Source”. In: *2008 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2008* (Las Vegas, NV, Mar. 31–Apr. 4, 2008). 2008, pp. 1657–1660.
doi:10.1109/ICASSP.2008.4517945.

[165] Peter Meerwald and Andreas Uhl. “Additive Spread-Spectrum Watermark Detection in Demosaicked Images”. In: *MM&Sec’09, Proceedings of the Multimedia and Security Workshop 2009* (Princeton, NJ, Sept. 7–8, 2009). 2009, pp. 25–32.
doi:10.1145/1597817.1597823.

[166] Daniele Menon and Giancarlo Calvagno. “Color Image Demosaicking: An Overview”. In: *Signal Processing: Image Communication* 26.8–9 (Oct. 2011), pp. 518–533.
doi:10.1016/j.image.2011.04.003.

[167] Aravind K. Mikkilineni, Nitin Khanna, and Edward J. Delp. “Texture Based Attacks on Intrinsic Signature Based Printer Identification”. In: *Media Forensics and Security II* (San Jose, CA, Jan. 18–20, 2010). Ed. by Nasir D. Memon, Jana Dittmann, Adnan M. Alattar, and Edward J. Delp. Vol. 7541. Proceedings of SPIE. Bellingham, WA: SPIE Press, 2010, 75410T.
doi:10.1117/12.845377.

[168] Steven J. Murdoch and Maximillian Dornseif. “Hidden Data in Internet Published Documents”. In: *Proceedings of the 21st Chaos Communication Congress* (Berlin, Germany, Dec. 27–29, 2004). 2004.
http://md.hudora.de/presentations/forensics/HiddenData-21c3.pdf.

[169] Gopal Narayanan and Yun Qing Shi. “A Statistical Model for Quantized AC Block DCT Coefficients in JPEG Compression and its Application to Detecting Potential Compression History in Bitmap Images”. In: *Digital Watermarking, 9th International Workshop, IWDW 2010. Revised Selected Papers* (Seoul, Korea, Oct. 1–3, 2010). Ed. by Hyoung-Joong Kim, Yun Qing Shi, and Mauro Barni. Vol. 6526. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Verlag, 2011, pp. 75–89.

doi:10.1007/978-3-642-18405-5_7.

[170] Ramesh Neelamani, Ricardo de Queiroz, Zhigang Fan, Sanjeeb Dash, and Richard G. Baraniuk. "JPEG Compression History Estimation for Color Images". In: *IEEE Transactions on Image Processing* 15.6 (June 2006), pp. 1365–1378.
doi:10.1109/TIP.2005.864171.

[171] John W. van Ness. "Dimensionality and Classification Performance with Independent Coordinates". In: *IEEE Transactions on Systems, Man and Cybernetics* 7.7 (July 1977), pp. 560–564.
doi:10.1109/TSMC.1977.4309771.

[172] Jerzy Neyman and Egon S. Pearson. "On the Problem of the Most Efficient Tests of Statistical Hypotheses". In: *Philosophical Transactions of the Royal Society A* 231 (1933), pp. 289–337.
doi:10.1098/rsta.1933.0009.

[173] Tian-Tsong Ng. "Statistical and Geometric Methods for Passive-blind Image Forensics". PhD Thesis. New York, NY: Graduate School of Arts and Sciences, Columbia University, 2007.
http://www.ee.columbia.edu/dvmm/publications/PhD_theses/ttng_thesis.pdf.

[174] Tian-Tsong Ng. "Camera Response Function Signature for Digital Forensics — Part II: Signature Extraction". In: *First IEEE International Workshop on Information Forensics and Security, WIFS 2009* (London, UK, Dec. 6–9, 2009). 2009, pp. 161–165.
doi:10.1109/WIFS.2009.5386461.

[175] Tian-Tsong Ng and Shih-Fu Chang. *A Data Set of Authentic and Spliced Image Blocks*. Tech. rep. ADVENT 203-2004-3. New York, NY: Department of Electrical Engineering, Columbia University, 2004.
http://www.ee.columbia.edu/dvmm/publications/04/TR_splicingDataSet_ttng.pdf.

[176] Tian-Tsong Ng and Shih-Fu Chang. "A Model for Image Splicing". In: *IEEE International Conference on Image Processing, ICIP 2004* (Singapore, Oct. 24–27, 2004). Vol. 2. 2004, pp. 1169–1172.
doi:10.1109/ICIP.2004.1419512.

[177] Tian-Tsong Ng and Shih-Fu Chang. *Classifying Photographic and Photorealistic Computer Graphic Images using Natural Image Statistics*. Tech. rep. ADVENT 220-2006-6. New York, NY: Department of Electrical Engineering, Columbia University, Oct. 2004.
http://www.ee.columbia.edu/dvmm/publications/04/ng-tech-report-nis-04.pdf.

[178] Tian-Tsong Ng and Mao-Pei Tsui. "Camera Response Function Signature for Digital Forensics — Part I: Theory and Data Selection". In: *First IEEE International Workshop on Information Forensics and Security, WIFS 2009* (London, UK, Dec. 6–9, 2009). 2009, pp. 156–160.
doi:10.1109/WIFS.2009.5386464.

[179] Tian-Tsong Ng, Shih-Fu Chang, and Qibin Sun. "Blind Detection of Photomontage using High Order Statistics". In: *IEEE International Symposium on Circuits and Systems, ISCAS 2004* (Vancouver, BC, May 23–26, 2004). Vol. 5. 2004, pp. V–688–691.
doi:10.1109/ISCAS.2004.1329901.

[180] Tian-Tsong Ng, Shih-Fu Chang, Jessie Hsu, Lexing Xie, and Mao-Pei Tsui. "Physics-Motivated Features for Distinguishing Photographic Images and Computer Graphics". In: *MULTIMEDIA '05, Proceedings of the 13th Annual ACM International Conference on Multimedia* (Singapore, Nov. 6–11, 2005). New York, NY: ACM Press, 2005, pp. 239–248.
doi:10.1145/1101149.1101192.

[181] Tian-Tsong Ng, Shih-Fu Chang, Ching-Yung Lin, and Qibin Sun. "Passive-blind Image Forensics". In: *Multimedia Security Technologies for Digital Rights*. Ed. by Wenjun Zeng, Heather Yu, and Ching-Yung Lin. Academic Press, 2006, chap. 15, pp. 383–412.

[182] Tian-Tsong Ng, Shih-Fu Chang, and Mao-Pei Tsui. "Using Geometry Invariants For Camera Response Function Estimation". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007* (Minneapolis, MN, June 17–22, 2007). 2007.
doi:10.1109/CVPR.2007.383000.

[183] Mike Nizza and Patrick J. Lyons. *In an Iranian Image, a Missile Too Many*. 2008.
http://thelede.blogs.nytimes.com/2008/07/10/in-an-iranian-image-a-missile-too-many/.

[184] Martin S. Olivier. "Using Sensor Dirt for Toolmark Analysis of Digital Photographs". In: *Advances in Digital Forensics IV, IFIP International Conference on Digital Forensics* (Kyoto, Japan, Jan. 27–30, 2008). Ed. by Indrajit Ray and Sujeet Shenoi. Vol. 285. IFIP International Federation for Information Processing. Boston, MA: Springer Verlag, 2008, chap. 16, pp. 193–206.
doi:10.1007/978-0-387-84927-0_16.

[185] Xunyu Pan and Siwei Lyu. "Region Duplication Detection Using Image Feature Matching". In: *IEEE Transactions on Information Forensics and Security* 5.4 (Dec. 2010), pp. 857–867.
doi:10.1109/TIFS.2010.2078506.

[186] Debra Parrish and Bridget Noonan. "Image Manipulation as Research Misconduct". In: *Science and Engineering Ethics* 15.2 (2008), pp. 161–167.
doi:10.1007/s11948-008-9108-z.

[187] Patrick Pérez, Michel Gangnet, and Andrew Blake. "Poisson Image Editing". In: *ACM Transactions on Graphics* 22.3 (July 2003): *Proceedings of ACM SIGGRAPH 2003*, pp. 313–318.
doi:10.1145/882262.882269.

[188] Tomáš Pevný and Jessica Fridrich. "Detection of Double-Compression in JPEG Images for Applications in Steganography". In: *IEEE Transactions on Information Forensics and Security* 3.2 (June 2008), pp. 247–258.
doi:10.1109/TIFS.2008.922456.

[189] Tomáš Pevný, Tomáš Filler, and Patrick Bas. "Using High-Dimensional Image Models to Perform Highly Undetectable Steganography". In: *Information Hiding, 12th International Conference, IH 2010. Revised Selected Papers* (Calgary, AB, June 28–30, 2010). Ed. by Rainer Böhme, Philip W. L. Fong, and Reihaneh Safavi-Naini. Vol. 6387. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 161–177.
doi:10.1007/978-3-642-16435-4_13.

[190] Peter W. Pfefferli. “Digitalfotografie in der Strafverfolgung [Digital Photography for Legal Purposes]”. In German. In: *Kriminalistik* 58.9 (Sept. 2004), pp. 573–577.

[191] Andreas Pfitzmann and Marit Hansen. *A Terminology for Talking about Privacy by Data Minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management*. (Version 0.34). Aug. 2010.
http://dud.inf.tu-dresden.de/Anon_Terminology.shtml.

[192] John C. Platt. “Probabilities for SV Machines”. In: *Advances in Large Margin Classifiers*. Ed. by Alexander J. Smola, Peter L. Bartlett, Bernhard Schölkopf, and Dale Schuurmans. MIT Press, 2000, chap. 5, pp. 61–74.
http://research.microsoft.com/pubs/69187/svmprob.ps.gz.

[193] Marie-Charlotte Poilpré, Patrick Perrot, and Hugues Talbot. “Image Tampering Detection Using Bayer Interpolation and JPEG Compression”. In: *First International Conference on Forensic Applications and Techniques in Telecommunications, Information, and Multimedia, e-Forensics 2008* (Adelaide, SA, Jan. 21–24, 2008). 2008.
http://portal.acm.org/citation.cfm?id=1363240.

[194] Alin C. Popescu. “Statistical Tools for Digital Image Forensics”. PhD Thesis. Hanover, NH: Department of Computer Science, Dartmouth College, 2005.
http://www.cs.dartmouth.edu/farid/publications/apthesis05.pdf.

[195] Alin C. Popescu and Hany Farid. *Exposing Digital Forgeries by Detecting Duplicated Image Regions*. Tech. rep. TR2004-515. Hanover, NH: Department of Computer Science, Dartmouth College, Aug. 2004.
http://www.cs.dartmouth.edu/reports/abstracts/TR2004-515.

[196] Alin C. Popescu and Hany Farid. “Statistical Tools for Digital Forensics”. In: *Information Hiding, 6th International Workshop, IH 2004. Revised Selected Papers* (Toronto, ON, May 23–25, 2004). Ed. by Jessica Fridrich. Vol. 3200. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer-Verlag, 2004, pp. 128–147.
doi:10.1007/978-3-540-30114-1_10.

[197] Alin C. Popescu and Hany Farid. “Exposing Digital Forgeries by Detecting Traces of Re-sampling”. In: *IEEE Transactions on Signal Processing* 53.2 (Feb. 2005), pp. 758–767.
doi:10.1109/TSP.2004.839932.

[198] Alin C. Popescu and Hany Farid. “Exposing Digital Forgeries in Color Filter Array Interpolated Images”. In: *IEEE Transactions on Signal Processing* 53.10 (Oct. 2005), pp. 3948–3959.
doi:10.1109/TSP.2005.855406.

[199] Rajeev Ramanath, Wesley Snyder, Youngjun Yoo, and Mark Drew. “Color Image Processing Pipeline”. In: *IEEE Signal Processing Magazine* 22.1 (Jan. 2005), pp. 34–43.
doi:10.1109/MSP.2005.1407713.

[200] Frank Ramsthaler, Mattias Kettner, Stefan Potente, Axel Gehl, Kerstin Kreutz, and Marcel A. Verhoff. “Original oder manipuliert? Authentizität und Integrität digitaler Bildmaterialien aus forensischer Sicht [Original or Manipulated? Authenticity and Integrity of Digital Images from a Forensic Perspective]”. In German. In: *Rechtsmedizin* 20.5 (Oct. 2010), pp. 385–392.

doi:10.1007/s00194-010-0669-1.

[201] Christian Riess and Elli Angelopoulou. “Scene Illumination as an Indicator of Image Manipulation”. In: *Information Hiding, 12th International Conference, IH 2010. Revised Selected Papers* (Calgary, AB, Canada, June 28–30, 2010). Ed. by Rainer Böhme, Philip W. L. Fong, and Reihaneh Safavi-Naini. Vol. 6387. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Verlag, 2010, pp. 66–80.
doi:10.1007/978-3-642-16435-4_6.

[202] Edward M. Robinson. *Crime Scene Photography*. 2nd ed. San Diego, CA: Academic Press, 2010.

[203] Alessia de Rosa, Francesca Uccheddu, Andrea Costanzo, Alessandro Piva, and Mauro Barni. “Exploring Image Dependencies: A New Challenge in Image Forensics”. In: *Media Forensics and Security II* (San Jose, CA, Jan. 18–20, 2010). Ed. by Nasir D. Memon, Jana Dittmann, Adnan M. Alattar, and Edward J. Delp. Vol. 7541. Proceedings of SPIE. Bellingham, WA: SPIE, 2010, 75410X.
doi:10.1117/12.840235.

[204] Kurt Rosenfeld and Husrev T. Sencar. “A Study of the Robustness of PRNU-based Camera Identification”. In: *Media Forensics and Security XI* (San Jose, CA, Jan. 19–21, 2009). Ed. by Edward J. Delp, Jana Dittmann, Nasir D. Memon, and Ping Wah Wong. Vol. 7254. Proceedings of SPIE. Bellingham, WA: SPIE, 2009, 72540M.
doi:10.1117/12.814705.

[205] Mike Rossner and Kenneth M. Yamada. “What’s in a Picture? The Temptation of Image Manipulation”. In: *Journal of Cell Biology* 166.1 (July 2004), pp. 11–15.
doi:10.1083/jcb.200406019.

[206] Seung-Jin Ryu, Hae-Yeoun Lee, Dong-Hyuck Im, Jung-Ho Choi, and Heung-Kyu Lee. “Electrophotographic Printer Identification by Halftone Texture Analysis”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2010* (Dallas, TX, Mar. 14–19, 2010). 2010, pp. 1846–1849.
doi:10.1109/ICASSP.2010.5495377.

[207] Frode Eika Sandnes. “Where Was that Photo Taken? Deriving Geographical Information from Image Collections Based on Temporal Exposure Attributes”. In: *Multimedia Systems* 16.4–5 (Aug. 2010), pp. 309–318.
doi:10.1007/s00530-010-0188-7.

[208] Yun Q. Shi, Chunhua Chen, and Wen Chen. “A Natural Image Model Approach to Splicing Detection”. In: *MM&Sec’07, Proceedings of the Multimedia and Security Workshop 2007* (Dallas, TX, Sept. 20–21, 2007). New York, NY: ACM Press, 2007, pp. 51–62.
doi:10.1145/1288869.1288878.

[209] Dmitry Sklyarov. *Forging Canon Original Decision Data*. Nov. 2010.
http://www.elcomsoft.com/canon.html.

[210] Matthew Sorell. “Conditions for Effective Detection and Identification of Primary Quantisation of Re-Quantized JPEG Images”. In: *International Journal of Digital Crime and Forensics* 1.2 (2009), pp. 13–27.
doi:10.4018/jdcf.2009040102.

[211] Matthew C. Stamm and K. J. Ray Liu. "Forensic Detection of Image Manipulation Using Statistical Intrinsic Fingerprints". In: *IEEE Transactions on Information Forensics and Security* 5.3 (Sept. 2010), pp. 492–506.
doi:10.1109/TIFS.2010.2053202.

[212] Matthew C. Stamm and K. J. Ray Liu. "Forensic Estimation and Reconstruction of a Contrast Enhancement Mapping". In: *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2010* (Dallas, TX, Mar. 14–19, 2010). 2010, pp. 1698–1701.
doi:10.1109/ICASSP.2010.5495488.

[213] Matthew C. Stamm and K. J. Ray Liu. "Anti-Forensics of Digital Image Compression". In: *IEEE Transactions on Information Forensics and Security* 6.3 (Sept. 2011), pp. 1050–1065.
doi:10.1109/TIFS.2011.2119314.

[214] Martin Steinebach, Mohamed El Ouariachi, Huajian Liu, and Stefan Katzenbeisser. "On the Reliability of Cell Phone Camera Fingerprint Recognition". In: *Digital Forensics and Cyber Crime, First International ICST Conference, ICDF2C 2009. Revised Selected Papers* (Albany, NY, Sept. 30–Oct. 2, 2009). Ed. by Sanjay Goel. Vol. 31. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 69–76.
doi:10.1007/978-3-642-11534-9_7.

[215] Guangling Sun, Zhoubiao Shen, and Yuejun Chen. "Color Filter Array Synthesis in Digital Image via Dictionary Re-demosaicing". In: *International Conference on Multimedia Information Networking and Security, MINES 2010* (Nanjing, China, Nov. 4–6, 2010). 2010, pp. 898–901.
doi:10.1109/MINES.2010.191.

[216] Sabine E. Süsstrunk and Stefan Winkler. "Color Image Quality on the Internet". In: *Internet Imaging V* (San Jose, CA, Jan. 19, 2004). Ed. by Simone Santini and Raimondo Schettini. Vol. 5304. Proceedings of SPIE. Bellingham, WA: SPIE, 2003, pp. 118–131.
doi:10.1117/12.537804.

[217] Ashwin Swaminathan. "Multimedia Forensic Analysis via Intrinsic and Extrinsic Fingerprints". PhD Thesis. Graduate School of the University of Maryland, College Park, MD, 2008.
http://hdl.handle.net/1903/8776.

[218] Ashwin Swaminathan, Min Wu, and K. J. Ray Liu. "Component Forensics of Digital Cameras: A Non-Intrusive Approach". In: *40th Annual Conference on Information Sciences and Systems* (Princeton, NJ, Mar. 22–24, 2006). 2006, pp. 1194–1199.
doi:10.1109/CISS.2006.286646.

[219] Ashwin Swaminathan, Min Wu, and K. J. Ray Liu. "A Component Estimation Framework for Information Forensics". In: *IEEE 9th Workshop on Multimedia Signal Processing, MMSP 2007* (Chania, Greece, Oct. 1–3, 2007). 2007, pp. 397–400.
doi:10.1109/MMSP.2007.4412900.

[220] Ashwin Swaminathan, Min Wu, and K. J. Ray Liu. "Nonintrusive Component Forensics of Visual Sensors Using Output Images". In: *IEEE Transactions on Information Forensics and Security* 2.1 (Mar. 2007), pp. 91–106.
doi:10.1109/TIFS.2006.890307.

[221] Ashwin Swaminathan, Min Wu, and K. J. Ray Liu. "Optimization of Input Pattern for Semi Non-Intrusive Component Forensics of Digital Cameras". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2007* (Honolulu, HI, Apr. 15–20, 2007). 2007, pp. II–225–228.
doi:10.1109/ICASSP.2007.366213.

[222] Ashwin Swaminathan, Min Wu, and K. J. Ray Liu. "A Pattern Classification Framework for Theoretical Analysis of Component Forensics". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2008* (Las Vegas, NV, Mar. 31–Apr. 4, 2008). 2008, pp. 1665–1668.
doi:10.1109/ICASSP.2008.4517947.

[223] Ashwin Swaminathan, Min Wu, and K. J. Ray Liu. "Digital Image Forensics via Intrinsic Fingerprints". In: *IEEE Transactions on Information Forensics and Security* 3.1 (Mar. 2008), pp. 101–117.
doi:10.1109/TIFS.2007.916010.

[224] Ashwin Swaminathan, Min Wu, and K. J. Ray Liu. "Component Forensics". In: *IEEE Signal Processing Magazine* 26.2 (2009): *Digital Forensics*, pp. 38–48.
doi:10.1109/MSP.2008.931076.

[225] Michael E. Tipping. "Sparse Bayesian Learning and the Relevance Vector Machine". In: *Journal of Machine Learning Research* 1 (June 2001), pp. 211–244.
http://www.jmlr.org/papers/volume1/tipping01a/tipping01a.pdf.

[226] Godfried T. Toussaint. "On the Divergence between two Distributions and the Probability of Misclassification of Several Decision Rules". In: *Proceedings of the Second International Joint Conference on Pattern Recognition* (Copenhagen, Denmark, Aug. 13–15, 1974). 1974, pp. 27–35.
http://www-cgrl.cs.mcgill.ca/~godfried/publications/divergence.pdf.

[227] Godfried T. Toussaint. "Probability of Error, Expected Divergence, and the Affinity of Several Distributions". In: *IEEE Transactions on Systems, Man and Cybernetics* 8.6 (June 1978), pp. 482–485.
doi:10.1109/TSMC.1978.4310001.

[228] G. Valenzise, V. Nobile, M. Tagliasacchi, and S. Tubaro. "Countering JPEG Anti-Forensics". In: *IEEE International Conference on Image Processing, ICIP 2011* (Brussels, Belgium, Sept. 11–14, 2011). 2011, pp. 1949–1952.
doi:10.1109/ICIP.2011.6115854.

[229] Lanh Tran Van, Sabu Emmanuel, and Mohan S. Kankanhalli. "Identifying Source Cell Phone using Chromatic Aberration". In: *IEEE International Conference on Multimedia and EXPO, ICME 2007* (Beijing, China, July 2–5, 2007). 2007, pp. 883–886.
doi:10.1109/ICME.2007.4284792.

[230] Ramarathnam Venkatesan, S.-M. Koon, Mariusz H. Jakubowski, and Pierre Moulin. "Robust Image Hashing". In: *IEEE International Conference on Image Processing, ICIP 2000* (Vancouver, BC, Sept. 10–13, 2000). Vol. 3. 2000, pp. 664–666. doi:10.1109/ICIP.2000.899541.

[231] Wei Wang, Jing Dong, and Tieniu Tan. "A Survey of Passive Image Tampering Detection". In: *Digital Watermarking, 8th International Workshop, IWDW 2009* (Guildford, UK, Aug. 24–26, 2009). Ed. by Anthony T. S. Ho, Yun Q. Shi, Hyoung Joong Kim, and Mauro Barni. Vol. 5703. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 308–322. doi:10.1007/978-3-642-03688-0_27.

[232] Weihong Wang and Hany Farid. "Exposing Digital Forgeries in Video by Detecting Double MPEG Compression". In: *MM&Sec'06, Proceedings of the Multimedia and Security Workshop 2006* (Geneva, Switzerland, Sept. 26–27, 2006). 2006, pp. 37–47. doi:10.1145/1161366.1161375.

[233] Weihong Wang and Hany Farid. "Exposing Digital Forgeries in Interlaced and Deinterlaced Video". In: *IEEE Transactions on Information Forensics and Security* 2.3 (Sept. 2007), pp. 438–449. doi:10.1109/TIFS.2007.902661.

[234] Zhou Wang and Alan C. Bovik. "Mean Squared Error: Love it or Leave it? A New Look at Signal Fidelity Measures". In: *IEEE Signal Processing Magazine* 26.1 (Jan. 2009), pp. 98–117. doi:10.1109/MSP.2008.930649.

[235] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. "Image Quality Assessment: From Error Visibility to Structural Similarity". In: *IEEE Transactions on Image Processing* 13.4 (Apr. 2004), pp. 600–612. doi:10.1109/TIP.2003.819861.

[236] Weimin Wei, Shuozhong Wang, and Zhenjun Tang. "Estimation of Rescaling Factor and Detection of Image Splicing". In: *11th IEEE International Conference on Communication Technology, ICCT 2008* (Hangzhou, China, Nov. 10–12, 2008). 2008, pp. 676–679. doi:10.1109/ICCT.2008.4716196.

[237] Jie Wu, Markad V. Kamath, and Skip Poehlman. "Detecting Differences Between Photographs and Computer Generated Images". In: *24th IASTED International Conference on Signal Processing, Pattern Recognition, and Applications* (Innsbruck, Austria, Feb. 15–17, 2006). 2006, pp. 268–273.

[238] Ting-Fan Wu, Chih-Jen Lin, and Ruby C. Weng. "Probability Estimates for Multi-class Classification by Pairwise Coupling". In: *Journal of Machine Learning Research* 5 (Aug. 2004), pp. 975–1005. http://www.jmlr.org/papers/volume5/wu04a/wu04a.pdf.

[239] Fuchun Xie, Teddy Furon, and Caroline Fontaine. "Towards Robust and Secure Watermarking". In: *MM&Sec'10, Proceedings of the 2010 ACM SIGMM Multimedia & Security Workshop* (Rome, Italy, Sept. 9–10, 2010). New York: ACM Press, 2010, pp. 153–159. doi:10.1145/1854229.1854258.

[240] Guanshuo Xu, Yun Qing Shi, and Wei Su. “Camera Brand and Model Identification Using Moments of 1-D and 2-D Characteristic Functions”. In: *IEEE International Conference on Image Processing, ICIP 2009* (Cairo, Egypt, Nov. 7–10, 2009). 2009, pp. 2917–2920.
doi:10.1109/ICIP.2009.5413341.

[241] Ido Yerushalmy and Hagit Hel-Or. “Digital Image Forgery Detection Based on Lens and Sensor Aberration”. In: *International Journal of Computer Vision* 92.1 (Mar. 2011), pp. 71–91.
doi:10.1007/s11263-010-0403-1.

[242] Hang Yu, Tian-Tsong Ng, and Qibin Sun. “Recaptured Photo Detection Using Specularity Distribution”. In: *IEEE International Conference on Image Processing, ICIP 2008* (San Diego, CA, Oct. 12–15, 2008). 2008, pp. 3140–3143.
doi:10.1109/ICIP.2008.4712461.

[243] Jun Yu, Scott A. Craver, and Enping Li. “Toward the Identification of DSLR Lenses by Chromatic Aberration”. In: *Media Watermarking, Security, and Forensics III* (San Francisco, CA, Jan. 24–26, 2011). Ed. by Nasir D. Memon, Jana Dittmann, Adnan M. Alattar, and Edward J. Delp. Vol. 7880. Proceedings of SPIE. Bellingham, WA: SPIE Press, 2011, 788010.
doi:10.1117/12.872681.