Counter-Forensics: Attacking Image Forensics

Rainer Böhme and Matthias Kirchner

PRE-PRINT VERSION. This chapter appears in:
H. T. Sencar and N. Memon (eds.): "Digital Image Forensics", © Springer, 2012.

Abstract This chapter discusses counter-forensics, the art and science of impeding or misleading forensic analyses of digital images. Research on counter-forensics is motivated by the need to assess and improve the reliability of forensic methods in situations where intelligent adversaries make efforts to induce a certain outcome of forensic analyses. Counter-forensics is first defined in a formal decision-theoretic framework. This framework is then interpreted and extended to encompass the requirements to forensic analyses in practice, including a discussion of the notion of authenticity in the presence of legitimate processing, and the role of image models with regard to the epistemic underpinning of the forensic decision problem. A terminology is developed that distinguishes security from robustness properties, integrated from post-processing attacks, and targeted from universal attacks. This terminology is directly applied in a self-contained technical survey of counter-forensics against image forensics, notably techniques that suppress traces of image processing and techniques that synthesize traces of authenticity, including examples and brief evaluations. A discussion of relations to other domains of multimedia security and an overview of open research questions concludes the chapter.

1 Definition of Counter-Forensics

This final chapter changes the perspective. It is devoted to digital image *counter*-forensics, the art and science of impeding and misleading forensic analyses of digital images.

Rainer Böhme
Westfälische Wilhelms-Universität Münster, Dept. of Information Systems, Leonardo-Campus 3, 48149 Münster, Germany e-mail: rainer.boehme@wi.uni-muenster.de

Matthias Kirchner
International Computer Science Institute, 1947 Center Street, Berkeley, CA 94704
e-mail: kirchner@icsi.berkeley.edu

Digital image forensics has become a hot topic over the past couple of years. Initially academic schemes found applications in domains like law enforcement, intelligence, private investigations, and media. Passive image forensics has promised to reestablish trust in digital images, which otherwise were deemed too easy to manipulate. But what stops perpetrators, spies and swindlers, who make efforts to manipulate images for their own profit anyway, from finding out forensic investigators' latest tricks and techniques? Then they can use this knowledge to cover up traces or—even worse—plant misleading traces.

Many forensic algorithms were not designed with such behavior in mind, hence they are easy to deceive. To justify the additional trust we place in digital images through forensics, it is important that the limits of forensics are known and will eventually be overcome. The only alternative would be a closed infrastructure of trustworthy acquisition devices that authenticate images actively [15]. This vision is certainly costly and most likely politically unviable. The other apparent solution of keeping passive forensic techniques secret is determined to fail. Its advantages are only temporary and highly uncertain [23]. Not to mention that the chain of causality leading to a conviction in a public court must withstand scrutiny of independent experts, who would learn the techniques and potentially compromise their secrecy.

The research field that challenges digital forensics and systematically explores its limitations against intelligent 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. Counter-forensics stands in a equally productive relation to forensics like cryptanalysis to cryptography. Therefore we borrow from the cryptanalysis terminology and call a counter-forensic scheme *attack* (against forensics). This reflects the strategic intention of the counterfeiter's action.

Besides the need to assess and improve the reliability of forensic methods, two more reasons motivate research on counter-forensics. First, many forensic techniques link images to the circumstances of their acquisition, e. g., the acquisition device or time. This indirectly reveals information about identities of the author or depicted subjects. This is not always desired. Researchers have studied systems providing *unlinkability* and *anonymity* in digital communications for some time [37]. All these efforts are useless if the link 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* [43]. 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 imply a loss of image quality. Camera manufacturers, for instance, compete on quality. It is questionable if they would sacrifice a competitive edge for making reverse engineering a little harder. Yet we are curious to see promising applications in the future.

The outline of this chapter is as follows. In the next section, we define counter-forensics formally in a general decision-theoretic framework. This framework is interpreted and extended to encompass the requirements to forensic analyses in practice in Section 3, including a discussion of the notion of authenticity in the presence of legitimate processing, and the role of image models with regard to the epistemic underpinning of the forensic decision problem. Section 4 develops a terminology that distinguishes security from robustness properties, integrated from post-processing attacks, and targeted from universal attacks. Section 5 contains a technical survey of counter-forensics against image forensics. The survey is structured by techniques that suppress traces of image processing and techniques that synthesize traces of authenticity. It includes examples of each technique as well as a brief evaluation. Section 6 adds a discussion of relations to other domains of multimedia security, notably steganography and robust digital watermarking. We conclude this chapter with an overview of open research questions in the final Section 7.

2 Theory

Digital image forensics refers to the collection of scientific methods to systematically infer particulars of an unknown image generation process from a given (set of) digital image(s). For a formal treatment of image forensics and counter-forensics, we have to define this generation process (Section 2.1), then describe forensic inference as a decision problem (Section 2.2), before counter-forensics can be introduced as means to influence the outcome of such decisions (Section 2.3).

2.1 Image Generation

Generation of natural images links the real world with digital representations thereof.

Definition 1. The *image generation function* $\mathsf{generate} : \mathscr{N} \times \boldsymbol{\Theta} \to \mathscr{I}$ maps observations of the infinite set of all conceivable natural phenomena $\mathcal{N} \in \mathscr{N}$ to the finite set of digitized images $\mathbf{I} \in \mathscr{I}$. This mapping is parametrized with a collection of parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

The parameters include, inter alia, the perspective, the time of the acquisition, the choice of the acquisition device, and its configuration (e. g., settings, lenses). It is convenient to understand the image generation process as a combination of both the image acquisition with a digital imaging device and subsequent post-processing.

Definition 2. Function $\mathsf{generate}$ is composed of a concatenation of an *image acquisition function* $\mathsf{acquire} \in \mathscr{A} : \mathscr{N} \to \mathscr{I}$ and an *image processing function* $\mathsf{process} \in \mathscr{P} : \mathscr{I}^+ \to \mathscr{I}$, where $\mathsf{acquire}$ and $\mathsf{process}$ are elements of the respective families of functions $\mathscr{A}$ of all possible image acquisition methods and $\mathscr{P}$ of

all possible image processing operations. The exact composition is defined by the parameters θ of generate.

In the above definition, operator $^+$ is the 'Kleene plus', which for a given set $\mathscr{I}$ is defined as $\mathscr{I}^+ = \bigcup_{n=1}^{\infty} \mathscr{I}^n$. Hence, function process may take an arbitrary positive number of digital images as input. The block diagram in Figure 1 illustrates the concept of image generation suggested by Definitions 1 and 2 (ignoring the possibility of multiple inputs to function process for the sake of simplicity). The case of *passive–blind* image forensics implies that the forensic investigator has no influence on generate (passive) and her knowledge about generate is limited to the information given in the image under analysis (blind).

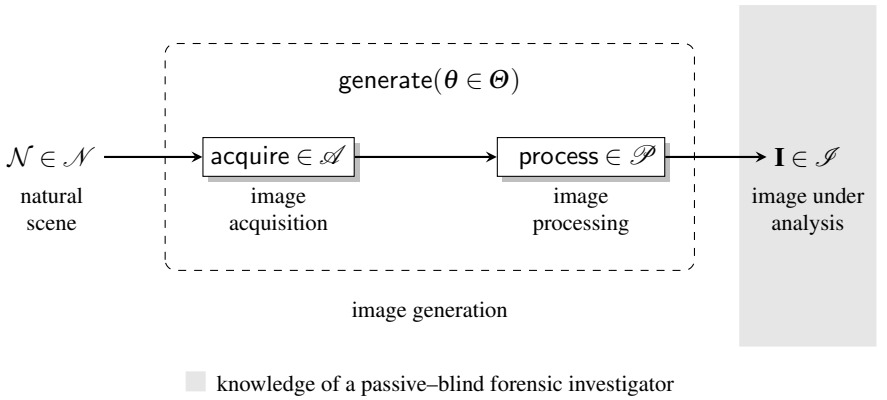


Fig. 1 General image generation function in the context of passive–blind image forensics

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 $\mathscr{P}$ thus explicitly includes the identity function, $\perp_{\mathscr{P}}: \mathbf{I} \mapsto \mathbf{I}$, i. e., no post-processing.

Definition 3. All image generation functions $(\mathsf{acquire}, \perp_{\mathscr{P}}) \in \mathscr{A} \times \mathscr{P}$ produce *original images* as opposed to *processed images* that result from generation functions $(\mathsf{acquire}, \mathsf{process}) \in \mathscr{A} \times \mathscr{P} \setminus \{\perp_{\mathscr{P}}\}$.

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.

Definition 4. All image generation functions $(\mathsf{acquire}, \mathsf{process}) \in \mathscr{A} \setminus \{\perp_{\mathscr{A}}\} \times \mathscr{P}$ produce *natural images* as opposed to *computer-generated images* that result from generation functions $(\perp_{\mathscr{A}}, \mathsf{process}) \in \mathscr{A} \times \mathscr{P}$. By definition, computer-generated images are processed images.

A further important attribute with regard to the image generation process is the notion of authenticity. A digital image $\mathbf{I}$ is called authentic if it is a valid projection of the natural phenomenon $\mathcal{N}$. Instances of process may impair authenticity.

To give a formal definition of authenticity, we note that the projection of one particular natural phenomenon $\mathcal{N}$ to an authentic image is not necessarily unique. There may exist many different mappings that yield *semantically equivalent* images. This means each element in a set of many different images $\mathbf{I}_1 \neq \mathbf{I}_2 \neq \cdots \neq \mathbf{I}_n$ is a valid representation of the same realization of nature $\mathcal{N}$. For example, in many cases it makes no difference with which digital camera a given event is captured, but each camera will produce a slightly different image. 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 the tuple $(\mathbf{I}, \theta, \mathcal{N})$ where $\mathcal{N}$ must be the realization of $\mathscr{N}$ under parameters θ.

Intuitively, also the *semantic meaning* of an image refers to the link between a depicted scene and the corresponding natural phenomenon. Yet this is more difficult to formalize, as the association of semantic meaning requires interpretation and it is highly context-dependent in general. We work around this difficulty and assume that semantic equivalence is measurable between images.

Definition 5. Two images $\mathbf{I}$ and $\mathbf{J} \in \mathscr{I}$ are *semantically equivalent* if there exists $\mathcal{N} \in \mathscr{N}$ such that

$$|\mathsf{dist}(\mathbf{I}, \mathcal{N}) - \mathsf{dist}(\mathbf{J}, \mathcal{N})| < d,$$

where $\mathsf{dist} : \mathscr{I} \times \mathscr{N} \to \mathbb{R}_+$ is a measure of the *semantic distance* between an image and a—real or imaginary—natural phenomenon, and d is a given threshold.

The *semantic resolution* is the ability of function dist to differentiate between very similar natural phenomena for a fixed image $\mathbf{I}$. This resolution depends on the *quality* of an image, or, more precisely, on the information conveyed in an image $\mathbf{I}$ about $\mathcal{N}$. Threshold d has to be chosen commensurate with the semantic resolution of the image with the lowest quality.

Equipped with the notion of semantic equivalence, we can finally define what qualifies an image as authentic.

Definition 6. All original natural images are authentic. Furthermore, for a given authentic image $\mathbf{I} = \mathsf{generate}(\mathcal{N}, \theta)$, a processed version $\mathbf{J} = \mathsf{process}(\mathbf{I})$ is called authentic if $\mathbf{I}$ and $\mathbf{J}$ are semantically equivalent with respect to $\mathcal{N}$.

Definitions 3 and 6 reflect the subtle yet significant difference between processed and counterfeit images. While each non-trivial instance of process destroys the originality of an image, it not necessarily impairs its authenticity. Whether or not a processed image will be considered as counterfeit ultimately depends on a given context and established habits. We further point out that computer-generated images are not counterfeits 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.

Figure 2 gives a schematic overview of the different classifications of digital images discussed so far. Each of the three layers depicts a different partition of the same function space $\mathscr{A} \times \mathscr{P}$. 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.

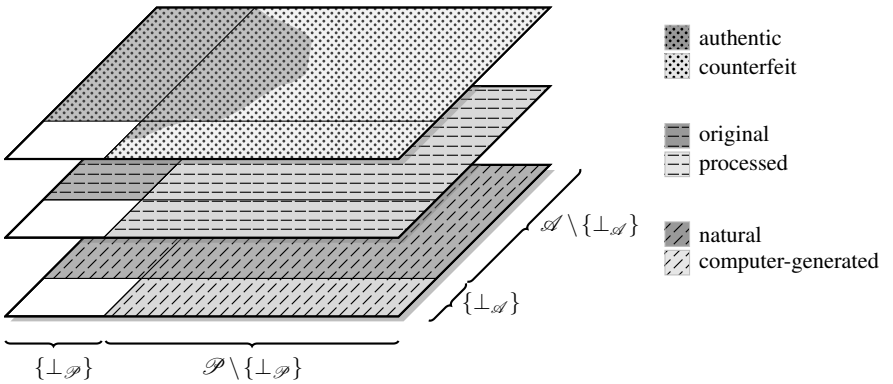


Fig. 2 Function space $\mathscr{A} \times \mathscr{P}$ of conceivable combinations of image acquisition and processing functions along with different classifications of the resulting digital images

2.2 Digital Image Forensics as a Classification Problem

The ultimate objective of passive–blind image forensics is to infer the authenticity of a given image without knowledge about the inputs of generate, notably the scene $\mathcal{N}$ and the parameters θ. Yet authenticity is often too hard to prove, so that forensic investigations resort to the inference on the inputs to generate. These serve as indicators which can be combined with side-information from other sources to make statements about the authenticity. Forensic analyses are possible if the functions acquire and process leave *identifying traces* in the resulting images. These traces can be used to distinguish between samples from different generation processes. Hence, digital image forensics is best described as a classification problem.

2.2.1 Classes in Digital Image Forensics

Forensic investigators define classes $\mathcal{C}_0, \ldots, \mathcal{C}_k$ to encapsulate parameter ranges of the generation function. The choice of the class space, denoted by $\mathscr{C}$, depends on the concrete application. For example, manipulation detection is usually stated as binary classification problem, $|\mathscr{C}| = 2$, with one class $\mathcal{C}_0$ for original images and another class $\mathcal{C}_1$ for processed images. In the case of source identification, the classes represent different imaging sources, e. g., different digital camera models or individual devices (typically $|\mathscr{C}| \gg 2$).

Definition 7. 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 $(\mathscr{A} \times \mathscr{P})_{(\not\mathcal{C})}$ so that all images $\mathbf{I}_{(\mathcal{C})}$ generated by $(\mathsf{acquire}, \mathsf{process}) \in (\mathscr{A} \times \mathscr{P})_{(\mathcal{C})}$ share common identifying traces.

Convention. To keep notations simple, we use $\mathbf{I}_{(k)}$ equivalent for $\mathbf{I}_{(\mathcal{C}_k)}$ when referring to instances of images of a particular class $\mathcal{C}_k \in \mathscr{C}$. Moreover, we write $\mathbf{I}_{(0)}$ for

authentic images and $\mathbf{I}_{(1)}$ for counterfeits whenever the context prevents ambiguities and the class space contains only these two classes.

Definitions 1–7 allow us to express various kinds of image forensics in a unified formal framework, as illustrated by the following examples.

Example 1. Natural versus computer-generated images: Class $\mathcal{C}_0$ of natural images contains all instance of images $\mathbf{I}_{(0)}$ generated by functions in the subspace $(\mathscr{A} \setminus \{\perp_{\mathscr{A}}\} \times \mathscr{P})$. Class $\mathcal{C}_1$ of computer-generated images entails all instance of images $\mathbf{I}_{(1)}$ generated by functions in the subspace $(\{\perp_{\mathscr{A}}\} \times \mathscr{P})$.

Example 2. Manipulation detection: Class $\mathcal{C}_0$ of original images contains all instances of images $\mathbf{I}_{(0)}$ generated by functions in the subspace $(\mathscr{A} \times \{\perp_{\mathscr{P}}\})$. Class $\mathcal{C}_1$ of processed images entails all instances of images generated by functions in the subspace $(\mathscr{A} \times \mathscr{P} \setminus \{\perp_{\mathscr{P}}\})$.

Example 3. Source identification via sensor noise: Class $\mathcal{C}_k$ of acquisition with sensor $k = 1, \ldots$ contains all instances of images $\mathbf{I}_{(k)}$ generated by functions in the subspace $(\mathscr{A}_k \times \mathscr{P})$ where $\mathscr{A}_k \subset \mathscr{A}$ is the set of all image acquisition functions of sensor k and $\bigcap_k \mathscr{A}_k = \emptyset$.

Example 4. (Ideal) temporal forensics: Class $\mathcal{C}_{t_1,t_2}$ of images acquired in the time interval $t_1 < t < t_2$ contains all instances of images $\mathbf{I}_{(\mathcal{C}_{t_1,t_2})}$ 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. In practice, temporal forensics today requires prior knowledge of the acquisition device so that more specific partitions have to be defined.

Note that classes are intentionally defined by partitioning the image generation process, not the image space $\mathscr{I}$. This is why in the examples above, we refer to *instances* of images, i. e., outputs of specific invocations of generate. A given image $\mathbf{I} \in \mathscr{I}$ with unknown provenance may be the result of different generation functions spanning more than one class. To resolve this ambiguity in the possibilistic framework, it is useful to take a probabilistic perspective.

Definition 8. Function $\mathcal{P}_{\mathcal{C}} : \mathscr{I} \to [0,1]$ is the likelihood function returning the conditional probability $\Pr(\mathbf{I}|\mathcal{C})$ of observing image $\mathbf{I}$ if the generation process falls in the partition of class $\mathcal{C}$. The probability reflects the empirical distributions of $\mathcal{N} \sim \mathscr{N}$ and $(\mathsf{acquire}, \mathsf{process}) \sim (\mathscr{A} \times \mathscr{P})_{(\mathcal{C})}$.

This probabilistic perspective allows us to quantify the ambiguity and derive decision rules, which on average minimize forensic decision errors.

2.2.2 Decision Rules

Given a class space $\mathscr{C}$, $|\mathscr{C}| \geq 2$, and an observed digital image $\mathbf{I}$ with unknown class, the forensic investigator needs a decision rule to assign $\mathbf{I}$ to a class $\mathcal{C}_*$.

Definition 9. A *digital image forensics algorithm* is given by a function $\mathsf{decide} : \mathscr{I} \to \mathscr{C}$ that assigns an image $\mathbf{I} \in \mathscr{I}$ to a class $\mathcal{C} \in \mathscr{C}$.

This decision rule now partitions the image space into disjoint classes, $\mathscr{I} = \bigcup_k \mathscr{R}_k$, such that all elements within a *decision region* $\mathscr{R}_k$ are assigned to class $\mathcal{C}_k$,

$$\mathscr{R}_k = \{\mathbf{I} \in \mathscr{I} \mid \mathsf{decide}(\mathbf{I}) = \mathcal{C}_k\}\,. \tag{1}$$

It is reasonable to assume that decisions are based on the the class probabilities conditional to the observed image, $\Pr(\mathcal{C}_k \,|\, \mathbf{I})$, which can be calculated from the likelihood functions in Definition 8 by using Bayes' theorem,

$$\Pr(\mathcal{C}_k \,|\, \mathbf{I}) = \frac{\Pr(\mathbf{I} \,|\, \mathcal{C}_k) \cdot \Pr(\mathcal{C}_k)}{\sum_i \Pr(\mathbf{I} \,|\, \mathcal{C}_i) \cdot \Pr(\mathcal{C}_i)} = \frac{\mathcal{P}_{\mathcal{C}_k}(\mathbf{I}) \cdot \Pr(\mathcal{C}_k)}{\sum_i \mathcal{P}_{\mathcal{C}_i}(\mathbf{I}) \cdot \Pr(\mathcal{C}_i)}\,. \tag{2}$$

In general, the larger is $\Pr(\mathcal{C}_k \,|\, \mathbf{I})$, the more evidence exists that $\mathbf{I}$ was generated by a function $(\mathsf{acquire}, \mathsf{process}) \in (\mathscr{A} \times \mathscr{P})_{(\mathcal{C}_k)}$. The concrete transformation of posterior probabilities into decisions depends on the algorithm decide and its decision rule. Many image forensics algorithms adhere to the minimum probability of error principle and decide for the class that maximizes $\Pr(\mathcal{C}_k \,|\, \mathbf{I})$ [2],

$$\mathsf{decide}(\mathbf{I}) = \mathcal{C}_* \quad \Leftrightarrow \quad \mathcal{C}_* = \arg\max_{\mathcal{C}_i \in \mathscr{C}} \Pr(\mathcal{C}_i \,|\, \mathbf{I})\,. \tag{3}$$

It is possible to impose additional constraints to ensure a reliable decision, for example by requiring a minimum a-posteriori probability,

$$\Pr(\mathcal{C}_* \,|\, \mathbf{I}) \geq p_{\min}\,, \tag{4}$$

or a minimum separability from the second-most probable class,

$$\Pr(\mathcal{C}_* \,|\, \mathbf{I}) - \max_{\mathcal{C}_i \in \mathscr{C} \setminus \mathcal{C}^*} \Pr(\mathcal{C}_i \,|\, \mathbf{I}) \geq p_{\text{sep}}\,. \tag{5}$$

Here, function decide has to return a special value for undecidable cases.

In many forensic problems, for which the class space is defined to comprise only two classes $\mathcal{C}_0$ and $\mathcal{C}_1$ (see for instance Examples 1 and 2), the decision problem can be expressed as a simple hypothesis test between

- H_0: the image under analysis $\mathbf{I}_{(k)}$ is generated by a process belonging to class $k = \mathcal{C}_0$, and
- H_1: the image under analysis $\mathbf{I}_{(k)}$ is generated by another process $k = \mathcal{C}_1$.

Because $\Pr(\mathcal{C}_0 \,|\, \mathbf{I}) + \Pr(\mathcal{C}_1 \,|\, \mathbf{I}) = 1$, and according to the theory of hypothesis tests, the optimal decision rule is given by the likelihood-ratio test,

$$\mathsf{decide}(\mathbf{I}) = \mathcal{C}_k \quad \Leftrightarrow \quad \frac{\mathcal{P}_{\mathcal{C}_k}(\mathbf{I})}{\mathcal{P}_{\mathcal{C}_{|k-1|}}(\mathbf{I})} > \tau\,, \qquad k \in \{0, 1\}\,. \tag{6}$$

2.3 Counter-Forensics

For a given image $\mathbf{I} = \mathbf{I}_{(k)}$, counter-forensics aims at preventing the assignment to the image's class $\mathcal{C}_k$. By suppressing or counterfeiting identifying traces, the counterfeiter creates a *counterfeit* $\mathbf{J} = \mathbf{J}_{(\hat{l})}$ with the intention to let it appear like an authentic member of an alternative class $\mathcal{C}_l \in \mathscr{C}$, $l \neq k$, when presented to the forensic investigator's function decide.

Convention. We use the subscript notation $_{(\hat{l})}$ to denote the intended class change of the counterfeit.

Definition 10. A digital image forensics algorithm decide is *vulnerable* to a *counter-forensic attack* if for a given image $\mathbf{I} = \mathsf{generate}(\theta)$

$$\exists\, \mathsf{attack} \in \mathscr{P},\ \mathbf{J} = \mathsf{attack}(\mathbf{I}) \quad \text{so that} \quad \mathsf{decide}(\mathbf{J}) \neq \mathsf{decide}(\mathbf{I})$$

subject to the constraints

1. $\mathbf{I}$ and $\mathbf{J}$ are semantically equivalent (*semantic constraint*), and
2. the probability of finding attack for a given $\mathbf{I}$ is not negligible within a given complexity bound (*computational constraint*).

The following examples illustrate how this definition matches existing counter-forensic strategies.

Example 5. A typical counter-forensic image manipulation of an authentic image $\mathbf{I}_{(0)}$ will involve two steps, first a transformation $\mathbf{I}_{(0)} \mapsto \mathbf{I}'_{(1)}$ which changes the semantic meaning according to the counterfeiter's intention, and second a counter-forensic attack $\mathbf{I}'_{(1)} \mapsto \mathbf{I}'_{(\hat{0})}$ to pretend authenticity of the counterfeit, i. e., $\mathsf{decide}\left(\mathbf{I}'_{(\hat{0})}\right) = \mathcal{C}_0$. Images $\mathbf{I}'_{(1)}$ and $\mathbf{I}'_{(\hat{0})}$ are semantically equivalent.

Example 6. Counterfeiting the source of an authentic image $\mathbf{I}$ involves a single application of a counter-forensic attack $\mathbf{I} \mapsto \mathbf{I}'$, possibly with the additional requirement that a specific target class $\mathcal{C}_{\text{target}} \stackrel{!}{=} \mathsf{decide}(\mathbf{I}') \neq \mathsf{decide}(\mathbf{I})$ is pretended.

Note that counterfeiting a particular target class generally needs to address both the suppression of identifying traces of the original class and synthesis of artificial traces of the target class. For example in PRNU-based digital camera identification [13], inserting the reference noise pattern of a target camera may lead to decisions for the new class $\mathcal{C}_{\text{target}}$. But the (distorted) fingerprint of the true camera is still present. A thorough forensic investigator may find abnormally high likelihood values $\mathcal{P}_{\mathcal{C}_k}\left(\mathbf{I}'_{(k)}\right), k \neq \mathcal{C}_{\text{target}}$, suspicious.

This leads us to the notion of *reliability* of a counter-forensic attack $\mathbf{I}_{(k)} \mapsto \mathbf{J}_{(\hat{l})}$ against all possible decision rules on a given class space $\mathscr{C}$. (Unlike Definition 10, which is formulated for a specific choice of function decide.) 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 $\mathcal{C}'_0$ and $\mathcal{C}'_1$ to represent the set of target (i. e., admissible) generation functions and attacks, respectively:

$$(\mathscr{A} \times \mathscr{P})_{(\mathcal{C}'_0)} \subseteq (\mathscr{A} \times \mathscr{P})_{(\mathcal{C}_k)} \tag{7}$$

$$(\mathscr{A} \times \mathscr{P})_{(\mathcal{C}'_1)} = (\mathscr{A} \times \mathscr{P})_{(\mathcal{C}_k)} \times \{\mathsf{attack}\}\,. \tag{8}$$

The concrete definition of class $\mathcal{C}'_0$ 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 $\mathcal{C}_k$) or to a particular class $\mathcal{C}_{\text{target}}$ (for instance to pretend a specific source device, cf. Example 6).

Careful counterfeiters in general strive to design their attacks so that samples of both classes $\mathcal{C}'_0$ and $\mathcal{C}'_1$ are indistinguishable. The decidability of the hypothesis test in Eq. (6) for all realizations of $\mathbf{I} \in \mathscr{I}$ can be measured by the Kullback–Leibler divergence between the two conditional probability distributions,

$$\mathsf{D}_{\mathrm{KL}}\left(\mathcal{P}_{\mathcal{C}'_0}, \mathcal{P}_{\mathcal{C}'_1}\right) = \sum_{\mathbf{I} \in \mathscr{I}} \mathcal{P}_{\mathcal{C}'_0}(\mathbf{I}) \log \frac{\mathcal{P}_{\mathcal{C}'_0}(\mathbf{I})}{\mathcal{P}_{\mathcal{C}'_1}(\mathbf{I})}\,. \tag{9}$$

Definition 11. A counter-forensic attack is ε-*reliable* against all digital image forensics algorithms on a class space $\mathscr{C}$ if for each pair $(\mathcal{C}'_0, \mathcal{C}'_1) \in \mathscr{C}' \times \mathscr{C}'$, $\mathbf{I}_{(\mathcal{C}'_k)} \sim \mathcal{P}_{\mathcal{C}'_k}$,

$$\mathsf{D}_{\mathrm{KL}}\left(\mathcal{P}_{\mathcal{C}'_0}, \mathcal{P}_{\mathcal{C}'_1}\right) \leq \varepsilon\,.$$

The unit of ε is bits or nats, depending on the base of the logarithm in Eq. (9). For the special case $\varepsilon = 0$, authentic and counterfeit images are drawn from the same distribution and the forensic investigator cannot gain any information from the analysis of $\mathbf{I}$. Hence, the counter-forensic attack is called *perfectly reliable*.

Note the similarity between Definition 11 and the notion of ε-*secure*, respectively *perfect steganography* in [5, 14]. Also in image forensics, ε bounds the error rates of the forensic investigator from below by the binary entropy function via the deterministic processing theorem. Further similarities between image forensics, counter-forensics, and other fields in information hiding are discussed in more detail in Section 6 below.

3 Practical Considerations

The theory in Section 2 provides a framework general enough to discuss a wide range of questions regarding digital image forensics and counter-forensics. However, only a few of the definitions are directly applicable in practice.

3.1 Epistemic Bounds

The main difficulty in applying the above equations is the lack of knowledge about the conditional probability distributions $\mathcal{P}_{\mathcal{C}}(\mathbf{I})$, which are given only empirically (cf. Def. 8). According to widely accepted epistemological paradigms, 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 [3, 4]. This is also reflected in Definition 1, which states that the support of $\mathscr{N}$ is infinite. Even after the transformation to the finite space $\mathscr{I}$, in general the support of $\mathcal{P}_{\mathcal{C}}(\mathbf{I})$ is too large and too heterogeneous to efficiently estimate distributions by sampling.

Even if we ignore this for a moment and assume that authentic images can efficiently be sampled, there remains the difficulty of sampling counterfeit images. Generating good counterfeits is a time-consuming manual task that largely depends on the counterfeiters' creativity. And it highly depends on the original image. This process is very hard to automate. The high cost of sampling is also reflected by the size and quality of available datasets that have been compiled in controlled environments for the purpose of image forensics research. Typical counterfeits are obtained by copying patches from within the same or other images, without sophisticated post-processing and without adaptivity to the depicted scene [35, 19]. Only few databases provide more realistic forgeries [8, 9], however without resolving the general trade-off between quality and quantity.

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 may be stated explicitly or—more often—implicitly. Models can be seen as a dimensionality reduction by projecting the high-dimensional image space $\mathscr{I}$ to a much smaller and more tractable subspace, e. g., a scalar in $\mathbb{R}$, on which decide is defined as a *discriminant function*.

Modeling images in low-dimensional model domains 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. Conversely, this also implies that the more restricted a forensic investigator's model of digital images is, the easier a counterfeiter can 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.

Counterfeiters are subject to the same limitations. They can never gain ultimate knowledge whether their image model is good enough so that no decision function can discriminate between two classes of authentic and counterfeit images. The theo-

retic reliability in Definition 11 cannot be calculated in the absence of knowledge of $\mathcal{P}_{\mathcal{C}_0}$ and $\mathcal{P}_{\mathcal{C}_1}$. A counter-forensic attack will only be successful as long as image forensics algorithms have not been refined accordingly. The interaction of forensics and counter-forensics can therefore be framed as competition for the best image model.

3.3 Blurred Notion of Authenticity

Another obstacle is that authenticity is hard to evaluate in practice. Although Definition 6 is convenient for a formal approach, many shades of gray may exist in practical situations (and need to be reflected in appropriate definitions of the class space $\mathscr{C}$).

3.3.1 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 the function acquire with imaging devices and then strictly apply Definitions 3 and 6: all original images captured with these devices are authentic. However, considering the sophistication of modern imaging devices, the situation is not as easy. Increasingly, post-processing and image enhancement become integral parts to the *internal* imaging pipeline. A forensic investigator has to accept that such processing inevitably raises the uncertainty of forensic decisions because device-internal processing and post-processing can generate very similar results. For example, many forensic techniques are sensitive to heavy device-internal quantization. While quantization in most cases preserves the semantic meaning of an image, it causes information loss by definition. This can make it harder to distinguish authentic from counterfeit images. The situation is even more complex when consumers are in the position to actively modify and extend the firmware of their devices.[1]

3.3.2 Analog Hole

The convention in Definition 6 to assume authenticity for every image generated with function process equal to the identity function does not reflect the possibility of image manipulations in or via the analog domain. While one can argue that detecting posed scenes in the real world is beyond the scope of digital image forensics, this limitation also excludes attacks which involve the reproduction and reacquisition of digitally processed images. In these attacks, and all iterations thereof, elements of $\mathscr{A}$ become part of $\mathscr{P}$. This blurs the distinction between the empirical functions

[1] 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`.

$\mathsf{acquire} \in \mathscr{A}$ and deterministic functions $\mathsf{process} \in \mathscr{P}$ further. To be very precise, one would have to extend our theory and define a new set of functions for digital-to-analog transformations. Then, pairs of this functions and acquire would have to be included into the transitive hull of $\mathscr{P}$. For the sake of brevity and clarity, we refrain from introducing new terminology and leave it with drawing the reader's attention on this blind spot of our theory.

3.3.3 Legitimate Post-Processing

While it is tempting to deny authenticity to every processed image *per se*, this simplification is too narrow for many realistic applications. Instead, there may be a subset $\mathscr{P}_{\text{legitimate}} \subset \mathscr{P}$ of legitimate processing operations which do not impair the authenticity of an image (cf. Figure 2).

This subset 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. Glossy magazines, by contrast, may not permit any quality reduction when acquiring images from photographers. Instead they may allow some basic color enhancement operations. There exist special codes of conduct, which specify what is considered legitimate post-processing for scientific journals [36, 11].

So a realistic model of authenticity will contain at least three categories, namely

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

The context in a practical situation defines whether the first two categories or just the first category shall be considered as authentic.

Example 7. 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, a party in doubt may present (or demand) the original raw files. The claim "these are the original raw files" 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 likelihood function $\mathcal{P}_{\mathcal{C}}(\mathbf{I}\,|\,\text{claim})$. This way, contextual knowledge can be incorporated in the formal framework and sharpen the notion of plausibility with probability distributions. Again, for brevity we refrain from extending our terminology in this chapter.

4 Classification of Counter-Forensic Techniques

Counter-forensic techniques against passive–blind image forensics can be classified along three dimensions. First, counterfeiters can generally exploit robustness or security weaknesses to mislead forensic analyses. Second, integrated and post-processing attacks vary in their position in the image generation process. Third, targeted and universal attacks differ in the (range of) attacked forensic algorithms. Figure 3 illustrates our classification and names representative examples for each relevant category. The following subsections discuss each of the dimensions in more detail.

4.1 Robustness versus Security

The distinction between legitimate and illegitimate post-processing becomes relevant for the distinction of robustness and security properties of forensic algorithms (see Section 3.3.3).

Definition 12. The *robustness* of a digital image forensics algorithms is defined by its reliability under legitimate post-processing.

In terms of counter-forensics, the lack of a clear separation between original and manipulated images increases the strategic options of a counterfeiter. If quality reduction, such as lossy compression or downscaling, is considered plausible and thus inconspicuous, a counterfeiter can eliminate subtle traces of illegitimate processing by subsequent quality reduction. Many known forensic algorithms are sensitive to strong quantization and fail to identify subtle traces in low-quality images. 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 [21].

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 $\mathscr{R}_{\mathcal{C}_1}$ entailing all manipulated images (subject to the constraints in Def. 10).

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. While this is a good indicator of the average reliability, it does not permit conclusions on the overall reliability. To complete the picture, also worst-case scenarios with sophisticated and intentional counterfeiters have to be considered. Resistance against *attacks* is directly associated with the *security* of forensic algorithms.

Definition 13. The *security* of a digital image forensics algorithm is defined by its reliability to detect intentionally concealed illegitimate post-processing.

In other words, security is the ability to withstand counter-forensics.

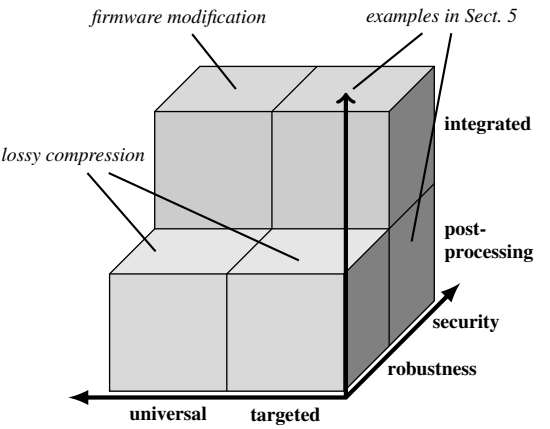


Fig. 3 Design space for counter-forensic techniques

Counterfeiters attacking security properties exploit specific knowledge about and shortcomings of the image model used by forensic investigators. The counterfeits are therefore purposely moved *in a particular direction* towards the decision boundary of the primary class (and just beyond). Such attacks are more powerful because their success does not depend on adjustable definitions of plausibility, but rather on weaknesses of forensic algorithms.

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

4.2 Integrated and Post-Processing Attacks

Post-processing attacks modify an image $\mathbf{I}_{(k)}$ such that the resulting counterfeit $\mathbf{I}_{(\hat{l})}$ does not reveal traces of the original class $\mathcal{C}_k$ anymore (cf. Example 5 in Section 2.3). Figure 4 illustrates that such attacks can be thought of as an additional processing step $\mathsf{attack} \in \mathscr{P}$ that supplements the original generation process $(\mathsf{acquire}, \mathsf{process}) \in (\mathscr{A} \times \mathscr{P})_{(\mathcal{C}_k)}$. Particular examples include lossy compression to take advantage of robustness issues, but also inverse flatfielding as a means to exploit specific weaknesses of digital camera identification based on sensor noise (cf. Section 5.2.2).

Integrated attacks, on the other hand, interact with or replace parts of the image generation process such that, instead of $\mathbf{I}_{(k)}$, the counterfeit is generated directly by a tuple $(\mathsf{acquire}', \mathsf{process}') \in (\mathscr{A} \times \mathscr{P})$. The modified functions $\mathsf{acquire}'$ and / or $\mathsf{process}'$ are specifically designed to avoid the formation of identifying traces or to mimic characteristics of the target class (see also Figure 4). In the aforementioned Example 5, an integrated attack would directly transform the original authentic image

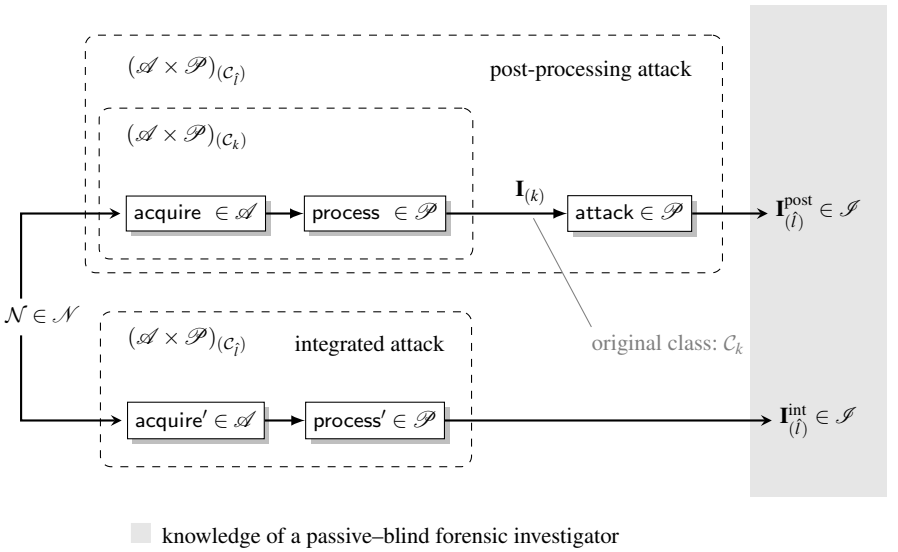


Fig. 4 Post-processing and integrated counter-forensic attacks. Post-processing attacks suppress and/or synthesize identifying traces subsequent to the original image generation process. Integrated attacks directly replace the original process with a counter-forensic variant

$\mathbf{I}_{(0)}$ to a semantically different counterfeit $\mathbf{I}'_{(\hat{0})}$ without ever releasing the detectable manipulation $\mathbf{I}'_{(1)}$. Note that this procedure is also covered by our formal description of counter-forensic attacks in Definition 10. It is always possible to express the (imaginary) map $\mathbf{I}'_{(1)} \mapsto \mathbf{I}'_{(\hat{0})}$ in terms of a post-processing function attack. Because integrated methods obviously require deep knowledge of the image generation process, they do not address robustness issues of forensic algorithms by definition. This is indicated in Figure 3, where the corresponding regions are left blank.

Integrated methods are relevant for manipulation detectors, where counterfeiters are hardly restricted in the choice of image processing primitives (or variants thereof). We will encounter several examples in Section 5. Integrated attacks to impede source identification are less obvious (apart from the pathological case of capturing a scene with a completely different device). 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 very powerful because device-specific traces need not leave the device at all.

4.3 Targeted and Universal Attacks

A further classification is borrowed from the context of steganalysis [14] and digital watermarking [10]. We call an attack *targeted*, if it exploits particulars and weak-

nesses of one specific forensic algorithm decide, which the counterfeiter usually knows. Such vulnerabilities directly relate to the image model implemented in the forensic algorithm, which leads to the notion of ε*-reliability with respect to a specific model*—yet another term that has is roots in steganography [14]. 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 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. In general, a mere combination of all known targeted counter-forensic techniques does not yield a universal attack; at least when typical low-dimensional image models are employed, which ignore the interference and interdependence of different attacks.

In the meantime, counterfeiters can exploit weak robustness and use lossy but legitimate processing whenever plausible. 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 true class, it very likely precludes claims about the originality of the corresponding image (cf. Section 3.3).

5 Selected Targeted Attacks

The literature on counter-forensic techniques is still very limited compared to the fast growth of publications on forensic techniques. In this section we survey the state of the art of image counter-forensics. The presentation is structured into techniques to suppress traces of possibly malicious image processing and techniques to restore or introduce artificial traces of seemingly authentic images.

5.1 Suppressing Traces of Image Processing

To suppress characteristic traces, variants of typical image processing operators were designed which destroy detectable structure by modulating the process with random noise. The methods differ in where in the process the randomness is introduced and how it is controlled to find a good trade-off between visual imperceptibility and undetectability by known forensic detectors.

5.1.1 Resampling

Resampling with interpolation is a very common image processing primitive which takes place during scaling, rotating, and shearing of images. For such affine coordinate

transformations, each pixel of the resampled image $\mathbf{I}'_{(1)}$ is a weighted sum of one or more pixels of the source image $\mathbf{I}_{(0)}$,

$$\mathbf{I}'_{(1)}(x,y) = \text{round}\left(\sum_i \sum_j \phi\left(\Delta\left(\mathbf{A}^{-1}\begin{pmatrix}x\\y\end{pmatrix}, \begin{pmatrix}i\\j\end{pmatrix}\right)\right)\mathbf{I}_{(0)}(i,j)\right), \tag{10}$$

where $\mathbf{A}$ is a 2×2 transformation matrix, $\Delta : \mathbb{R}^2 \times \mathbb{Z}^2 \to \mathbb{R}^+$ is a distance function that returns the distance between two coordinate vectors, and $\phi : \mathbb{R}^+ \to \mathbb{R}$ is the interpolation kernel.

The interpolation step ensures visually appealing outcomes when pixel intensities are mapped from discrete positions of a source grid to a discrete destination grid without a one-to-one correspondence between source and destination positions. Depending on the relative position of source and destination grid, systematic dependencies between pixels and their local neighborhood are introduced. In many cases, the strength of this dependence alternates periodically in space, which leads to identifiable resampling artifacts. This is so because function Δ takes only a few different values for many combinations of input coordinates. See Chapter 9 in this volume for a review of methods to measure and interpret these traces.

Most automatic resampling detectors exploit the periodicity of dependencies between pixels, largely because those traces can be added up over multiple periods, thereby reducing the influence of noise and other interfering factors like the image content. Therefore, to effectively suppress traces of resampling, any periodic structure must be avoided. Basic signal processing theory suggests interpolation with a sinc-kernel, which is theoretically optimal for downscaling [47]. However this kernel requires infinite support. Aside from computational demands, it is also impractical because of boundary effects in finite images.

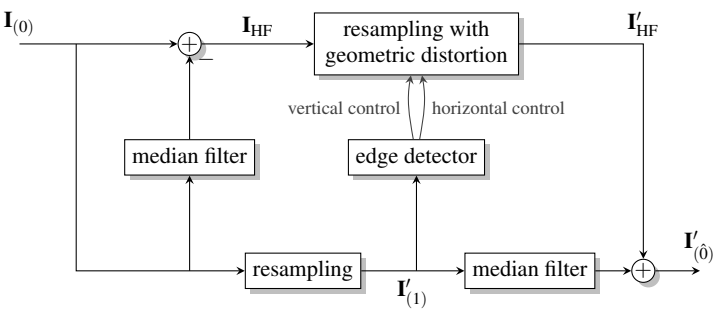


Fig. 5 Block diagram of undetectable resampling [25]

Another approach is to perturb the interpolation process with non-linear filters and noise [24, 25]. Both ideas complement each other in the so-called *dual-path attack*. This integrated attack decomposes the image into a low-frequency component and a high-frequency component as depicted in Figure 5. The high-frequency component

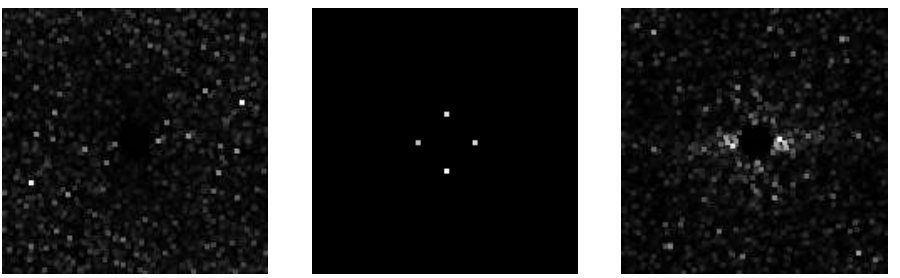
$\mathbf{I}_{\mathrm{HF}}$ of a source image $\mathbf{I}_{(0)}$ is obtained by subtracting the output of a median filter, a non-linear low-pass filter with windows size $2s+1$, from the source image:

$$\mathbf{I}_{\mathrm{HF}}(x,y) = \mathbf{I}_{(0)}(x,y) - \mathrm{median}\left\{\mathbf{I}_{(0)}(i,j) \mid \sup\left(|x-i|,|y-j|\right) \leq s\right\}. \tag{11}$$

This component is resampled by a modified function which adds noise to the real-valued destination positions before interpolation,

$$\mathbf{I}'_{\mathrm{HF}}(x,y) = \sum_i \sum_j \phi\left(\Delta\left(\mathbf{A}^{-1}\begin{pmatrix}x\\y\end{pmatrix} + \begin{pmatrix}n_{\mathrm{hor},x,y}\\n_{\mathrm{ver},x,y}\end{pmatrix}, \begin{pmatrix}i\\j\end{pmatrix}\right)\right)\mathbf{I}_{\mathrm{HF}}(i,j). \tag{12}$$

This procedure is called *geometric distortion*. Scalars n_{hor} and n_{ver} are real-valued displacements for horizontal, respectively vertical geometric distortion drawn independently for each pair (x,y) from a zero-mean Gaussian random source $N(0,\sigma)$. Moderate geometric distortion effectively reduces detectable periodic artefacts, but it suffers from the side-effect that visible jitter appears at sharp edges in the image. To prevent this, the degree of the geometric distortion, i. e., the variance of the random source, is attenuated by the output of an edge detector. For better results, this control can be implemented for horizontal and vertical edges independently. This ensures that a horizontal edge is not affected by visible vertical geometric distortion while at the same time avoiding that measurable periodicities appear along the horizontal direction, and vice versa.



(a) original (b) scaled to 105 % (c) scaled to 105 % with dual-path attack

Fig. 6 Resampling peaks in the FFT-transformed linear predictor residue of a grayscale image. The spectral images were enhanced with a maximum filter and scaled to the maximum contrast range for better printing quality (source: [25])

The low-frequency component is first resampled with conventional methods. This intermediate result serves as input to the edge detector. In a second step, a median filter is applied. This non-linear filter effectively removes detectable periodicities from resampled images. However, by its very nature, it removes high frequencies with a non-linear and non-trivial cutoff function. To minimize the visual impact, both frequency components are added together after independent resampling,

$$\mathbf{I}'_{(\hat{0})}(x,y) = \text{round}\left(\mathbf{I}'_{\text{HF}}(x,y) + \text{median}\left\{\mathbf{I}'_{(1)}(i,j) \mid \sup\left(|x-i|,|y-j|\right) \leq s\right\}\right). \tag{13}$$

Figure 6 demonstrates the effectivity of this method for a single example. All three images are Fourier-transforms of linear predictor residue. Figure 6 (a) is the reference for an untampered image. Figure 6 (b) shows the result for the same image after scaling to 105 % of the original size using conventional bilinear interpolation. The characteristic peaks are clearly visible. Finally, Figure 6 (c) displays the same analysis results after resampling with the described dual-path attack. The characteristic peaks have disappeared and the spectrum is dominated by scattered local maxima, similar (but not identical) to the spectrum of the original image.

Larger experiments on the detectability of this method are documented in [25]. The quantitative results from up to 200 images are summarized in Figures 7 and 8 for scaling and rotation, respectively. While rotation, upscaling and moderate downscaling is very well detectable when done in the conventional way, the dual-path attack can lower the detection rates substantially. Rotations of more than 45° are equally (un)detectable by symmetry. Nevertheless, there remains a residual risk for the counterfeiter since in 10–20 % of the cases, the processing is still detectable. We also suspect that the dual-path attack, while suppressing known traces of resampling, may leave new kinds of traces. Besides some specific works on the detectability of sole median filtering [27, 7, 51], we are not aware on any research on how detectable these traces are.

The authors of [25] also evaluated the quality loss of the dual-path attack and report peak signal-to-noise ratios (PSNR) between 30 and 40 dB for scaling (more distortion for downscaling) and about 40 dB for rotation independent of the angle. The distortion is measured between the attacked resampled image $\mathbf{I}'_{(\hat{0})}$ and the conventionally resampled version of the image $\mathbf{I}'_{(1)}$ for each image and then aggregated by taking the average distortion over 100 images.

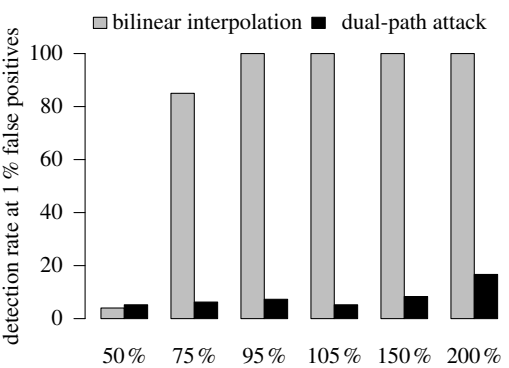


Fig. 7 Quantitative results for undetectable scaling: the dual-path attack reduces detectable traces of upscaling and moderate downscaling (experimental results of [25])

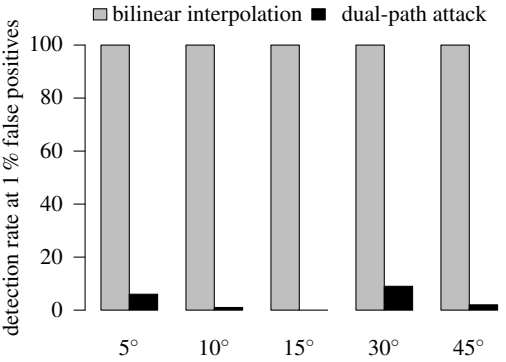


Fig. 8 Quantitative results for undetectable rotation: the dual-path attack reduces detectable traces of resampling after rotation (experimental results of [25])

5.1.2 Contrast Enhancement

Contrast enhancement is prone to leave characteristic traces in an image's histogram due to the non-linear mapping $f : \mathbb{Z} \to \mathbb{R}$ of integer values and subsequent rounding to integers,

$$\mathbf{I}'_{(1)}(i,j) = \text{round}\big(f\left(\mathbf{I}_{(0)}(i,j)\right)\big). \tag{14}$$

Typically function f increases monotonically. For gamma correction, it takes the form

$$f(x) = (2^{\ell}-1)\left(\frac{x}{2^{\ell}-1}\right)^{\gamma}, \tag{15}$$

where ℓ is the bit depth in bits and γ is the correction parameter. Values $\gamma < 1$ imply contrast reduction and values $\gamma > 1$ imply contrast enhancement.

Figure 9 visualizes the effect of Eq. (14) on the histogram of the modified image $\mathbf{I}'_{(1)}$. We distinguish two situations. First, not a single discrete value of the pre-image x is mapped after rounding to the value $f(x) = 3$ in the image. The corresponding histogram bin in $\mathbf{I}'_{(1)}$ remains empty, thus leading to a *gap* in the histogram. Second, multiple discrete values of the pre-image x are mapped after rounding to the value $f(x) = 8$ in the image. For typical pre-images with smooth histograms, this bin of the image's histogram receives a multiple of the hits of its adjacent bins, thus leading to a *peak* in the histogram. Forensic detectors can use the presence of gaps and peaks in a histogram as indications of prior processing with contrast enhancement [41].

An integrated attack against detectors which exploit these characteristics has been proposed in [6]. The idea is to identify histogram bins which are suspect of becoming a gap or peak. Then the result of the non-linear mapping is perturbed by adding noise to all pixels in the affected bins and their directly adjacent bins,

$$f^*(x) = f(x) + n \quad \text{with} \quad n \sim N(0, \sigma_x). \tag{16}$$

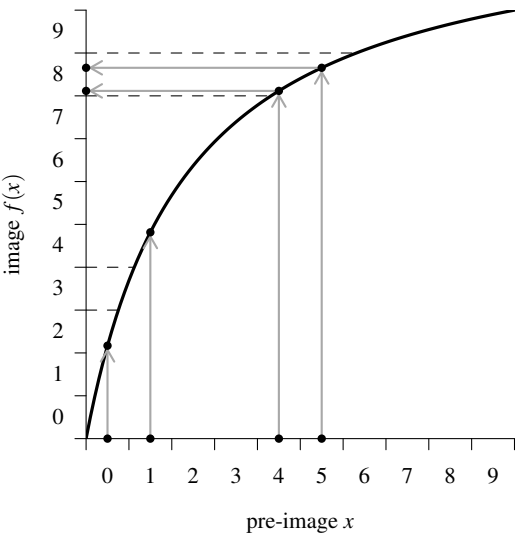


Fig. 9 Formation of gaps and peaks after rounding of non-linearly mapped integers; dashed lines indicate rounding margins

In principle, σ_x can be set depending on the proximity of a histogram bin to the next peak or gap. In practice, the authors of [6] report good results for a fixed dispersion parameter $\sigma = 1$ for all values of x. Figure 10 shows that this procedure effectively smoothes out the histogram of gamma-corrected versions of an example image (bottom and top rows), unlike conventional contrast enhancement (middle rows).

Figure 11 reports the results of a quantitative evaluation of this attack. The bars show detection rates of a detector which measures histogram peaks and gaps by identifying high-frequency components in the Fourier-transformed histogram [41]. The threshold has been set to allow 10 % false positives in a set of 693 natural images. Observe that the detector correctly detects all contrast enhancements regardless of the parameter choice. This is effectively prevented if the gamma correction is carried out with the integrated attack. The detection rates of the attacked images drop further to empirically zero if the number of tolerable false positives is set to 5 %. In practice, thresholds of less than 1 % might be required. For this experiment, Cao et al. [6] report PSNRs between the gamma-corrected image with attack and the gamma-corrected image without attack of about 47.5 dB. Some images deviate towards even higher values (i.e., relatively less distortion).

The attack can be extended to a post-processing attack by first estimating the positions of gaps and peaks in the histogram's frequency domain and then adding noise. This variant produces slightly inferior PSNR because the unrounded real values are not available and the noise has to be added to already rounded integers. Rounding errors and noise add up, whereas they partly cancel out in the integrated attack. Another potential problem are incorrect estimated of gaps and peaks, which can lead to inappropriate values of σ_x if σ is not fixed. We are not aware of any research on the practical relevance of this issue.

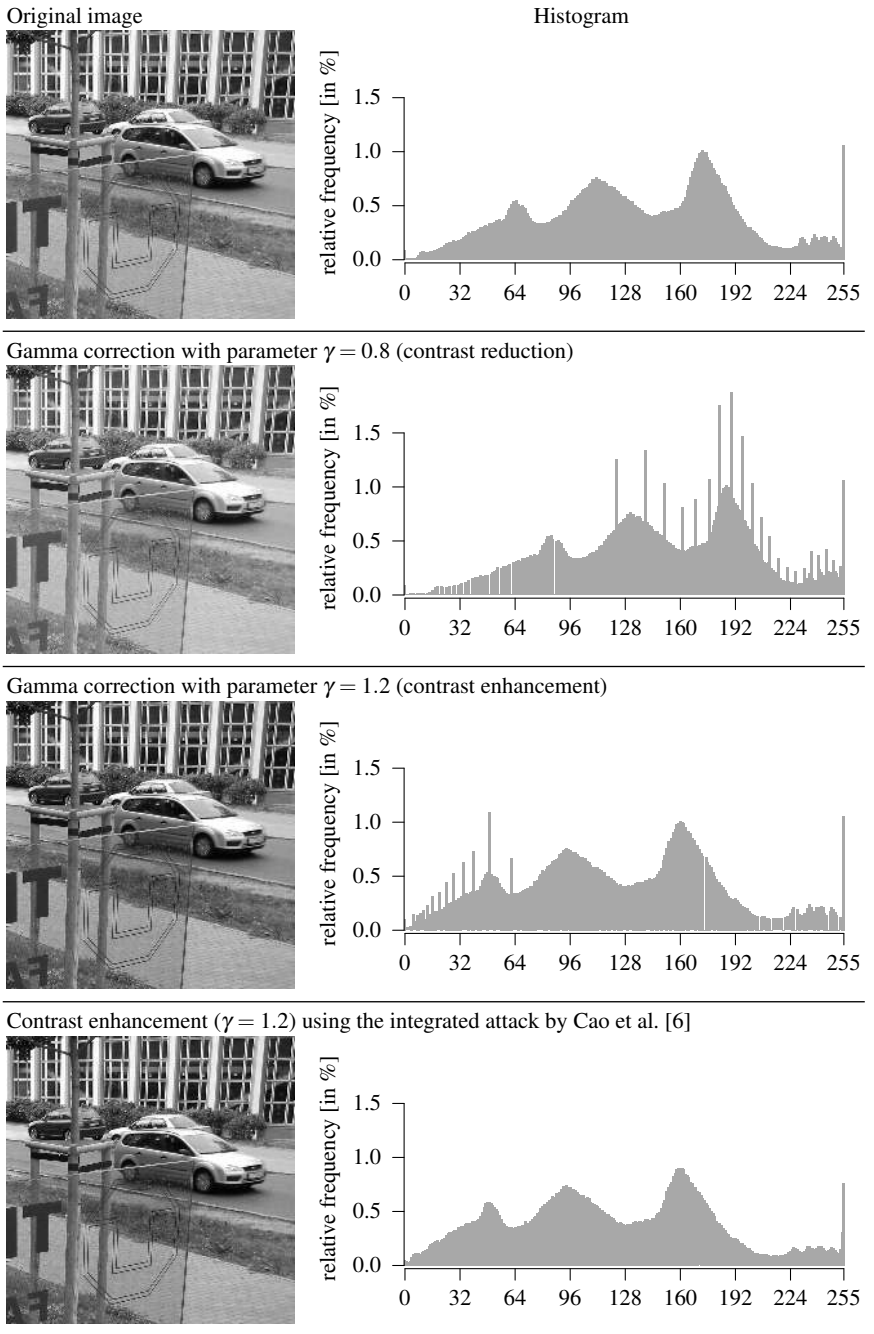


Fig. 10 Suppressing traces of double-quantization after non-linear mappings

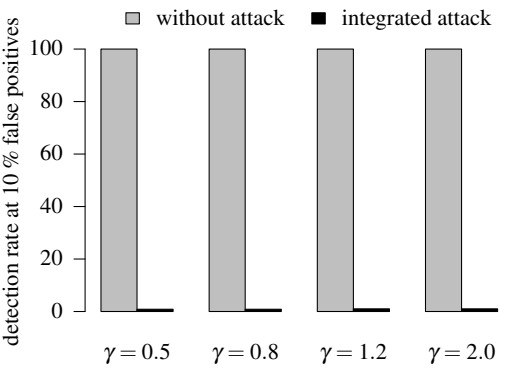


Fig. 11 Quantitative results for undetectable contrast enhancement: the integrated attack suppresses detectable traces in the histogram (aggregation of experimental results of [6])

5.1.3 Lossy Compression

A series of works [43, 44, 42] presents post-processing attacks to remove traces of lossy compression from a given image. This involves fixing two kinds of features which are exploited by current forensic detectors of the compression history: traces of dequantization and discontinuities at block boundaries. Specific counter-forensics have been proposed against detectors of each feature.

Hiding Traces of Dequantization

JPEG compression involves cutting a source image into blocks of 8×8 pixels. Each block is transformed to the frequency domain with the discrete cosine transformation (DCT). The resulting 64 real-valued coefficients are quantized with frequency-depedent quantization factors $q \geq 1$ before they are rounded to integers and finally stored with lossless Huffman encoding. Let y be the unquantized DCT coefficient, then the stored value $\tilde{y}$ is given by

$$\tilde{y} = \text{round}\left(\frac{y}{q}\right). \tag{17}$$

For a fixed frequency band, say DCT $(2,2)$ coefficients, the quantization factor q depends on the desired compression ratio and thus the retained quality. Higher values of q imply lower quality (and smaller file size) because the error between recovered coefficient value $\hat{y}$ and the original value y on average increases with q,

$$\begin{aligned} \mathrm{E}(|\hat{y}-y|) &= \mathrm{E}(|q\tilde{y}-y|) && (18)\\ &= \mathrm{E}\left(\left|q\cdot\text{round}\left(\frac{y}{q}\right)-y\right|\right) && (19)\\ &= \mathrm{E}(|q\cdot e|) = q\cdot\mathrm{E}(|e|) \qquad \text{and} \quad \mathrm{E}(|e|)>0. && (20) \end{aligned}$$

Now if a JPEG image is decompressed, the recovered coefficients $\hat{y}$ only take multiples of q. It is easy to detect a previous JPEG compression of any given image in spatial domain by applying block-wise DCT and checking the histograms of the DCT coefficients for gaps between multiples of a previous quantization factor. Figure 12 illustrates this and contrasts the smooth DCT histogram of a never-compressed image to the histogram of the same image after JPEG decompression with quality setting $Q = 60\,\%$. Observe the obvious peaks at multiples of the quantization factor $q = 10$. Note that unlike in the case of contrast enhancement, the gaps are not necessarily perfect in the sense that the actual frequency is zero. We rather observe "elephant feet" artifacts at the bottom of each histogram peak. This is so because during decompression, the real-valued intensity values from the inverse DCT are rounded to integers and truncated to the eligible value range. The rounding and truncation errors from all pixels in the block add up and may perturb the recalculated DCT coefficient beyond one rounding margin. Yet the errors are typically too small to smooth out the gaps entirely, even for $q = 2$.

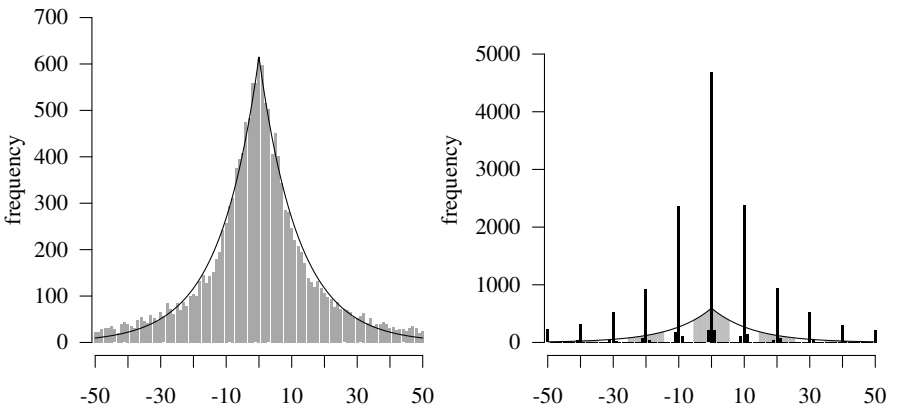


Fig. 12 Histograms of DCT $(2,2)$ coefficients for never-compressed (left) and JPEG-decompressed (right) image with fitted Laplace distribution superimposed. The JPEG quality is $Q = 60\,\%$, corresponding to a quantization factor of $q = 10$. Small pertubations from the theoretical values in the right histogram are from rounding errors in the spatial domain and subsequent transformation back into the DCT domain. Shaded areas in the right histogram indicate the value range over which the middle peak is distributed according to the Laplace model. Note the different scales

A post-processing attack to smooth out and restore such histograms is described in [43]. It uses the fact that DCT coefficients of natural images can be modeled reasonably well with a Laplace distribution [29]—except frequency band $(1,1)$, which requires special treatment. The good fit can be confirmed in the left histogram of Figure 12. The Laplace distribution is a symmetric one-parameter distribution with density function

$$f(x) = \frac{\lambda}{2} \exp(-\lambda \cdot |x|). \tag{21}$$

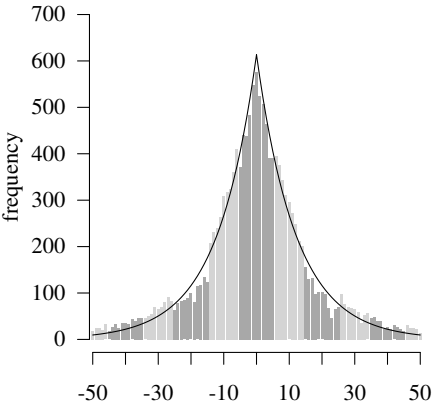


Fig. 13 Histogram of DCT $(2,2)$ coefficients after JPEG compression with quality $Q = 60\,\%$ and subsequent histogram restoration as proposed in [43]. The solid line is the fitted Laplace image model and the shaded bars indicate values which were expanded from the same quantized coefficient value. Compare to Figure 12

The authors of [43] employ a maximum likelihood estimator of parameter $\lambda > 0$ for the discretized case of the Laplace distribution. This model can be estimated from the peaks of a JPEG-decompressed histogram. In a second step, all coefficients are redistributed by adding noise according to conditional probability functions for each peak and each DCT frequency band. This way, the shape of a never-compressed DCT histogram is restored pretty well, as confirmed in the example of Figure 13. Since no parametric model is known for the DCT $(1,1)$ coefficients, noise following a uniform prior between two rounding margins is introduced. This is very similar to the post-processing variant of the attack against contrast enhancement detectors (see Section 5.1.2 above).

The effectiveness of this attack has been evaluated with quantitative experiments on 244 images, which were compressed using JPEG qualities $Q = 90\,\%$, $70\,\%$, and $50\,\%$. The restored images were presented to a state-of-the-art JPEG quantization matrix estimator [12]. An image was classified as pre-compressed if one or more quantization factors were estimated greater than one, i. e., $\hat{q} > 1$. After applying the post-processing attack, the compression history could be detected for only $4\,\%$ of the images pre-compressed with quality $Q = 90\,\%$, $7\,\%$ of the images pre-compressed with quality $Q = 70\,\%$, and $18\,\%$ of the images pre-compressed with quality $Q = 50\,\%$ [43]. The original publication does not report comparable detection rates for unattacked images, nor does it state the false positive rate. While the former are likely to reach almost $100\,\%$ for the chosen quality settings, we lack a good prior for the false positive rate. With regard to the retained image quality, only a single PSNR value is given at 41.6 dB. This example image has been pre-compressed with JPEG quality $Q = 60\,\%$ before restoration. We are not aware of more generalizable quality measurements. Note that recent publications call into question the perceptual quality [50] and the claimed statistical undetectability [28, 49] of this attack.

This attack is adapted to wavelet-based image compression in [42].

Suppression of Block Boundaries

Discontinuities at the boundaries of the 8×8 blocks are another indicative artifact of previous JPEG compression. Several researchers have investigated ways to remove blockiness artifacts from decompressed images, however mainly with the aim to increase the perceptual quality [31, 52]. The requirements in counter-forensics differ in that statistical undetectability has higher priority than visual experience. Therefore a post-processing attack against the blockiness detector of [12] is proposed in [44]. The attack uses a combination of a median filter and additive Gaussian noise,

$$\mathbf{I}'_{(\hat{0})}(x,y) = \text{round}\left(\text{median}\left\{\mathbf{I}'_{(1)}(i,j) \mid \sup(|x-i|,|y-i|) \leq s\right\} + n_{x,y}\right), \qquad (22)$$

where $n \sim N(0,\sigma)$ is drawn from a zero-mean Gaussian distribution independently for each pixel. The method is surprisingly effective, as can be seen in Figure 14. With moderately invasive parameter settings, $s = 1$ and $\sigma^2 = 2$, the attack outperforms the best-performing of the two visual blockiness reduction methods [31, 52] for JPEG qualities $Q = 70\,\%$ and above (hatched black bars in the figure). Stronger compression creates more pronounced block artifacts, which can be suppressed by adjusting the strength of the attack to $s = 2$ and $\sigma^2 = 3$. This way, even substantial and definitely noticeable JPEG compression can be effectively hidden from detectors which evaluate just a single blockiness criterion. For extreme settings like $Q = 10\,\%$, even the strong attack is not effective anymore. The original publication [44] does not provide any measures of retained image quality.

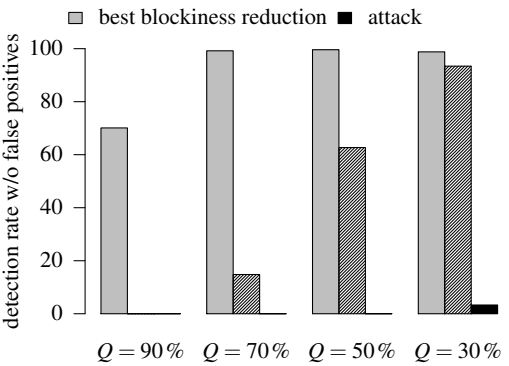


Fig. 14 Quantitative results for the suppression of JPEG blockiness artifacts: comparison of moderate ($s = 1, \sigma^2 = 2$, hatched bars) and strong ($s = 2, \sigma^2 = 3$, solid) counter-forensics against the best performing visual blockiness reduction (results from [44])

While the results against the specific detector in [12] are very clear, a note of caution is warranted. The attack has not been tested against a detector of median filtering [27]. While it is uncertain if this detector still works in the presence of additive noise and thus poses a serious threat, another concern arises on the choice of the blockiness detector. Prior efforts of suppressing even subtle JPEG blockiness after steganographic embedding in the DCT domain [40] has proven effective only against the very specific blockiness criterion used in the objective function of the

blockiness reduction algorithm. However, it turned out to be easily detectable as soon as the statistical criterion is replaced by an alternative one [48]. More research is needed to understand the relaibility of this attack against detectors based on different or a combination of several blockiness indicators.

5.2 Adding Traces of Authentic Images

Certain forensic techniques test the authenticity of images by verifying the integrity of traces of the common image acquisition or processing chain. Tampering very likely destroys these traces. In this case, the task of counter-forensics is to restore or synthesize the traces of authentic images. All attacks presented in this section are post-processing attacks.

5.2.1 Synthesis of Color Filter Array Pattern

Local dependencies between pixel values in different color channels of color images have attracted the interest of forensic investigators pretty early on [39]. These dependencies emerge from the way color images are captured by different sensors using color filter arrays (CFA pattern, see also Chapter 1 in this volume). More precisely, typical sensors capture a $n \times m$ matrix $\mathbf{I}$ of intensity values. The light received at each cell (i. e., pixel) went through one of three physical color filters: red, green, and blue. To obtain a full color image of the dimension $n \times m$, the missing color information of the two remaining color channels is filled in by interpolation,

$$\mathbf{I}_k = \mathsf{interpolate}(\mathsf{select}(\mathbf{I}, k)) \quad \text{for} \quad k \in \{\text{red}, \text{green}, \text{blue}\}. \tag{23}$$

Function $\mathsf{select} : \mathbb{Z}^{n \times m} \times \{\text{red}, \text{green}, \text{blue}\} \to \mathbb{Z}^{n' \times m'}$ isolates all pixels captured with a filter of a specific color. Function $\mathsf{interpolate} : \mathbb{Z}^{n' \times m'} \to \mathbb{Z}^{n \times m}$ expands the color channel to the original size by filling the missing values with interpolated values from adjacent pixels. (Note that $m' < m$ and $n' < n$.)

A simple method to restore the CFA pattern from a tampered color image $\mathbf{I}'_{(1)} = \left(\mathbf{I}'_{(1),\text{red}}, \mathbf{I}'_{(1),\text{green}}, \mathbf{I}'_{(1),\text{blue}}\right)$ is to straightly reapply the color filter interpolation [20],

$$\mathbf{I}'_{(\hat{0}),k} = \mathsf{interpolate}(\mathsf{select}(\mathbf{I}'_{(1),k}, k)) \quad \text{for} \quad k \in \{\text{red}, \text{green}, \text{blue}\}. \tag{24}$$

However, this procedure is far from optimal because it discards information in the unselected parts of each color channel, i. e., $\mathsf{select}(\mathbf{I}'_{(1),l}, k)$ with $k, l \in \{\text{red}, \text{green}, \text{blue}\}$ and $l \neq k$. To avoid this, the interpolation of each color channel can be formulated as a matrix product,

$$\vec{\mathbf{I}}_k = \mathbf{H}_k \vec{\mathbf{I}}. \tag{25}$$

Matrix $\mathbf{H}_k$ of dimension $nm \times nm$ holds the interpolation weights for color channel k. The notation $\vec{\mathbf{I}}$ denotes that matrices are vectorized to column vectors of length nm. Ignoring rounding errors for a moment, Eq. (25) should hold for authentic images whereas it is most likely violated after tampering. This can be expressed as a non-zero residual ε,

$$\vec{\mathbf{I}'}_{(1),k} = \mathbf{H}_k \vec{\mathbf{I}'} + \varepsilon. \tag{26}$$

The interpretation of this equation as a least-squares problem gives a method to synthesize CFA pattern with minimal distortion. An image $\mathbf{I}'$ that is compatible with $\mathbf{I}'_{(1)}$ and minimizes the L_2-norm $\|\varepsilon\|$ can be found by solving

$$\vec{\mathbf{I}'_k} = \left(\mathbf{H}_k^\mathsf{T} \mathbf{H}_k\right)^{-1} \mathbf{H}_k^\mathsf{T} \vec{\mathbf{I}'}_{(1),k}. \tag{27}$$

The general solution to Eq. (27) is computationally impractical due to the inversion of an $nm \times nm$ matrix. The solution in [26] exploits structure and sparsity of $\mathbf{H}$, which allows to derive linear time algorithms and approximations. Once $\mathbf{I}'$ is obtained, it can be inserted in Eq. (23) to generate the color channels $\mathbf{I}'_{(\hat{0}),k}$ which contain a seemingly authentic CFA pattern and exhibit minimal distortion compared to $\mathbf{I}'_{(1),k}$.

Note that the minimum is found under the assumption of continuous signals. Discretization and rounding errors may lead to a slight divergence from the optimal solution of the discrete optimization problem. Finding algorithms to (approximately) solve this problem efficiently remains an open research question.

(a) original (b) 9×9 median filtered (c) CFA synthesis after median filtering

Fig. 15 CFA peaks in the FFT-transformed linear predictor residue of a red color channel. The spectral images were enhanced with a maximum filter and scaled to the maximum contrast range for better printing quality (source: [26])

Figure 15 demonstrates this post-processing attack. The spectra are taken from linear predictor residue of the red color channel of one example image that has been taken as raw file from a digital camera. This is the method of choice for analyzing CFA pattern [39]. The four distinct peaks in Figure 15 (a) appear exactly where expected for the red channel. Hence they indicate authenticity. After applying a

median filter globally, the peaks disappear (Figure 15 (b)). We have chosen the median filter as a simple global operation, but other manipulation steps result in a similar effect. Note that invalid CFA pattern can be identified even if the violations appear only locally. Figure 15 (c) finally shows the reappearing peaks after CFA synthesis. This spectrum is visually indistinguishable from the one belonging to the original image.

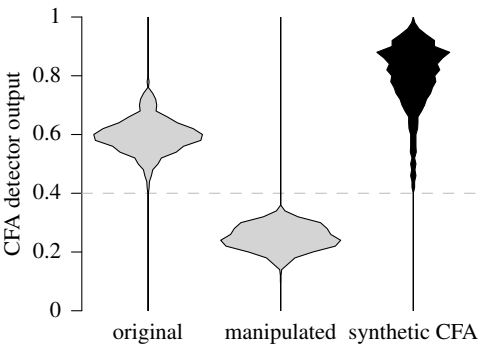


Fig. 16 Violin plots for quantitative results of a CFA synthesis experiment. Distribution of CFA peak detector outputs from red channels of 1000 images (data from [26])

Quantitative results with a peak detector come to a slightly different conclusion. Figure 16 displays the detector response distribution for three sets of 1000 images each. It is clearly visible that original and manipulated images can be separated perfectly with a decision threshold of $\tau = 0.4$. Moreover, the attack is successful in every single case. However the shape of the distributions for original and synthesized CFA pattern differ substantially. The shift of probability mass towards higher values indicates that synthesized CFA pattern appear even more authentic than original CFA pattern. This over-fitting can lead to suspicion. It is most likely caused by the fact that CFA interpolation is not the last step in a chain of device-internal processing operations and the synthesis method is focused on restoring the traces of a single operation. We are not aware of research on remedies. A first step could be to blend parts of $\mathbf{I}'_{(\hat{0})}$ and $\mathbf{I}'_{(1)}$ to shape the detector response distribution.

The retained quality has been measured for the same set of images. The median gain in PSNR between the straight method of Eq. (24) and the distortion minimization method of [26] has been reported at 1.2 dB for the red channel and 0.9 dB for the green channel. The difference is because function select discards less information for the green than for the red and blue channels. Therefore the quality gain for the blue channel is about as large as the gain for the red channel.

5.2.2 Substitution of Sensor Noise Pattern

A very powerful feature for forensic analyses are sensor defects in the image acquisition device (see Chapter 5 in this volume). Stable parts of the sensor noise, which

emerges from variations in the hardware manufacturing process and sensor wear-out, leave traces that are unique for each individual sensor. This noise pattern is useful to link (parts of) images with acquisition devices, for examples by extracting a sensor fingerprint from the photo response non-uniformity (PRNU).

Here we describe an attack against a predecessor of the image source identification method presented in Chapter 5. Instead of the peak-to-correlation energy measure for multiplicative sensor fingerprints, a simpler correlation detector was proposed in the seminal publication [32],

$$\mathcal{C}_* = \arg\max_{\mathcal{C}_k \in \mathscr{C}} \mathrm{cor}(\mathbf{I}, \hat{\mathbf{K}}_{\mathcal{C}_k}) \quad \text{with}$$
$$\mathrm{cor}(\mathbf{I}, \hat{\mathbf{K}}_{\mathcal{C}_k}) = \sum_x \sum_y \mathsf{norm}\left(\mathbf{I} - F(\mathbf{I})\right)(x,y) \cdot \mathsf{norm}\left(\hat{\mathbf{K}}_{\mathcal{C}_k}\right)(x,y). \tag{28}$$

Function $\mathsf{norm} : \mathbb{R}^{n\times m} \to \mathbb{R}^{n\times m}$ normalizes its argument by subtracting the mean and dividing by the variance. Function $F : \mathbb{R}^{n\times m} \to \mathbb{R}^{n\times m}$ is a low-pass filter to approximately separate image content from the noise residual. Matrix $\hat{\mathbf{K}}_{\mathcal{C}_k}$ is an estimated sensor noise pattern obtained from simple averaging of noise residuals $\sum_i \left(\mathbf{I}_{(k),i} - F\left(\mathbf{I}_{(k),i}\right)\right)$ of a set of doubtlessly authentic reference images acquired with sensor $\mathcal{C}_k$.

It is possible to mislead a detector based on Eq. (28) with the following procedure [17]. Suppose a given image $\mathbf{I}_{(k)}$ belonging to the class $\mathcal{C}_k$ shall be manipulated such that it is detected as belonging to the class $\mathcal{C}_l$, $l \neq k$, i. e., $\mathbf{I}_{(k)} \mapsto \mathbf{I}_{\left(\hat{l}\right)}$. Under the assumption that the counterfeiter has access to the primary acquisition device, she can produce a sufficiently large number L of

1. dark frames $\mathbf{I}^{\bullet}$ to estimate the dark current component of the sensor noise pattern $\hat{\mathbf{D}}_{(k)}$ by simple averaging, and
2. homogeneously illuminated frames $\mathbf{I}^{\circ}$ to estimate the flat-field frame $\hat{\mathbf{F}}_{(k)}$.

The flat-field frame $\mathbf{F}$ holds the PRNU component of the sensor pattern noise adjusted for the dark current component,

$$\hat{\mathbf{F}}_{(k)} = \frac{1}{L}\sum_{i=1}^{L}\left(\mathbf{I}^{\circ}_{(k),i} - \hat{\mathbf{D}}_{(k)}\right). \tag{29}$$

The tuple $\left(\hat{\mathbf{D}}, \hat{\mathbf{F}}\right)$ is a better representation of the sensor noise than the joint estimate $\hat{\mathbf{K}}$, which may contain parts of PRNU and dark current. Counterfeiting an images' origin now involves two steps, first the suppression of the original sensor noise pattern,

$$\mathbf{I}_{(\hat{1})} = \frac{\mathbf{I}_{(k)} - \hat{\mathbf{D}}_{(k)}}{\alpha \hat{\mathbf{F}}_{(k)}}, \tag{30}$$

where α is a scalar chosen to preserve the average luminance of the original image; second the insertion of a counterfeit sensor noise pattern,

$$\mathbf{I}_{(\hat{l})} = \alpha \mathbf{I}_{(\hat{\perp})} \hat{\mathbf{F}}_{(l)} + \hat{\mathbf{D}}_{(l)}. \tag{31}$$

If flat-field frame $\hat{\mathbf{F}}_{(l)}$ and dark frame $\hat{\mathbf{D}}_{(l)}$ are unavailable, for example because the counterfeiter has no access to the pretended acquisition device, an estimate $\hat{\mathbf{K}}_{(l)}$ from publicly available images can be used as substitute. However, this increases the risk of failure if the forensic investigator performs a triangle test with the same images (cf. [18] and Chapter 5 in this volume).

Figure 17 demonstrates how effectively the original noise pattern can be suppressed by the method of Eq. (30). The violin plot shows probability distributions of the correlation coefficient, Eq. (28), for 5×350 images acquired with 4 different devices. The reference pattern $\hat{\mathbf{K}}$ was estimated from images of the Canon S70 digital camera. As expected, only images taken with the Canon S70 produce significantly positive correlation coefficients. Observe that the probability mass of the black violin, representing Canon S70 images after suppression of the noise pattern, is almost as low as for the unrelated devices. In this example, the flat-field and dark frames were averaged from $L = 20$ images $\mathbf{I}^{\bullet}$ and $\mathbf{I}^{\circ}$.

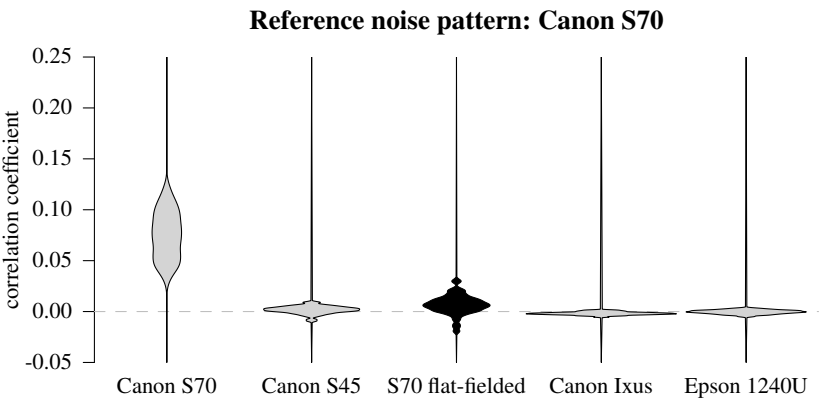


Fig. 17 Violin plots showing quantitative results of effective suppression of sensor noise pattern by flat-fielding (data of 350 images from [17])

Figure 18 continues the experiment and shows the results after a counterfeit noise pattern of the Canon S45 device has been inserted to the flat-fielded Canon S70 images using Eq. (31). Now both real and counterfeit Canon S45 images produce high correlation coefficients. Note that the shape of the distributions differ somewhat, though less than between different devices (compare Figure 17 and Figure 18). So in practice, a forensic investigator who has only a couple of images at her disposal, may have a hard time to extract useful information from the samples of different distributions.

The retained quality of this attack, if measured by PSNR between original $\mathbf{I}_{(k)}$ and counterfeited image $\mathbf{I}_{(\hat{l})}$, is below 30 dB on average. This seems notably worse than

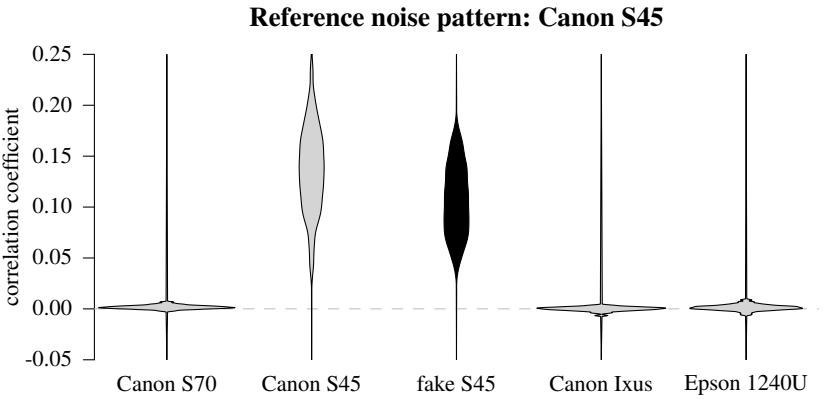


Fig. 18 Violin plots showing quantitative results for false classifications after insertion of a counterfeit S45 sensor noise pattern into images acquired with the Canon S70 camera (data from [17])

all other attacks surveyed in this chapter. A large part of this apparent degradation is due to the flat-fielding operation, which in fact is known as a technique to *improve* the visual quality of digital images. This improvement is later offset by the insertion of the counterfeit noise pattern, but both improvement and degradation are additive in the PSNR measure. So the overall quality loss is smaller than it appears.

This attack has been replicated several times, for example in [30], and in [46, 45] for cell-phone cameras. We are not aware of substantial refinements of this specific attack, for example adaptations to the peak-to-correlation energy detector.

The general potential weakness of PRNU-based image forensics to noise insertion attacks has already been mentioned in the seminal publication [32]. A defense strategy under the assumption that forensic investigator and counterfeiter share the images used for the attack is outlined in [33] and expanded and evaluated in [18].

5.2.3 Other Techniques

The above-described counter-forensic techniques to add traces of authentic images may need to be accompanied by additional adjustments. For example, the file structure and meta-data must be updated to be consistent with the pretended acquisition device. Due to the high number of mutually dependent specifications, this is not a trivial task. If the pretended acquisition device stores images as JPEG by default, then the JPEG quantization tables have to be adjusted to fit the manufacturer's internal table generation algorithm. A procedure to do this is to first suppress quantization artifacts (cf. Section 5.1.3) and subsequently recompress the images with quantization tables of a counterfeited acquisition device [44].

6 Relations to Steganography and Digital Watermarking

The previous sections introduced and discussed a formal framework for the description of counter-forensic attacks and surveyed the literature for concrete implementations. In this penultimate section, we broaden 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 [25]. 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.

Figure 19 illustrates how these disciplines relate to digital image forensics and counter-forensics. The figure 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 $\mathcal{C}_0$ to the class of stego images $\mathcal{C}_1$. 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 $\mathcal{P}_{\mathcal{C}_0}$ and $\mathcal{P}_{\mathcal{C}_1}$ (cf. Eq. (9)). 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 $\mathbf{I}_{(0)}$ have been modified, often with the aim to change its semantic meaning. The new semantic meaning of the counterfeit $\mathbf{I}'_{(1)}$ can be regarded as the 'message' to be transmitted. Unlike counter-forensics, steganography is not explicitly limited by perceptual constraints. Cover images are solely a means to communicate hidden messages. This leaves the steganographer more degree of freedom to choose the cover, and thus stego image.

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 with achievable detectability. The stronger a manipulating operation interferes with the inherent image structure, the harder it is to feign an authentic image.

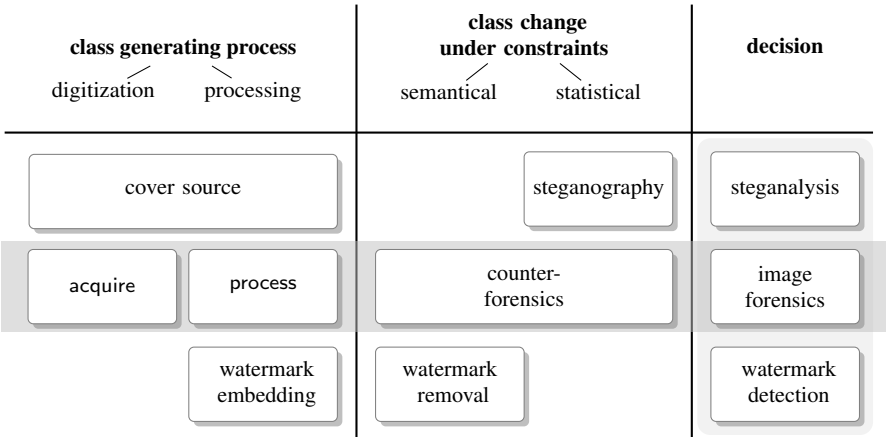


Fig. 19 Relation of image forensics and counter-forensics to other fields in information hiding

Another analogy exists between counter-forensics and *attacks against digital watermarking* schemes, where images of class $\mathcal{C}_1$ 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 $\mathcal{C}_0$, 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 [22], 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 4.2. 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).

Our brief excursion to related fields in information hiding highlights the strong overlap between counter-forensics, steganography and attacks against digital watermarking. While particularities of each field should not be disregarded, 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, where research on adversaries and attacks is a common practice and widely accepted. 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 [3, 1]. Moreover, the literature now reports digital watermarking schemes that directly interact with the image generation process [34], which suggests to employ

(counter-)forensic techniques as building blocks for the design of attacks against digital watermarks and the detection thereof. In summary, we see the relations in Figure 19 as an outline for a comprehensive theory of information hiding. Formalizing unified theoretical foundations seems a promising goal to advance the field.

7 Countering Counter-Forensics: An Open Challenge

As image forensics matures and finds applications in practice, its results may be used to rule on momentous, and sometimes controversial, decisions in the real world. It is unavoidable that actors who might be impaired by certain outcomes will try to influence the decisions towards a more favorable outcome for themselves. Consequently, forensic methods will be under attack. To avoid that the field as a whole loses credibility, image forensics has to anticipate the existence of counter-forensics and come up with new countermeasures [18].

This chapter has shown that current forensic techniques are vulnerable to relatively simple targeted attacks. Forensics investigators have several options to react to counter-forensics. Attacks against a lack of robustness can be warded off by demanding higher image quality standards (recall the courtroom example of Section 3.3.3). Fixing the security weaknesses fundamentally requires better image models. While it is generally hard to find good image models, finding models that are 'good enough' to withstand simple counter-forensics seems much more feasible.

As a starting point, single-criterion detectors can be replaced with detectors that use image models with multiple (but not many) dimensions. This makes it considerably harder for the counterfeiter to move an image into the right detection region *in each dimension at the same time*. If the number of dimensions grows unmanageable, machine-learning techniques may be used as discrimination functions. It is remarkable—and somewhat reassuring from the forensic investigator's point of view—that so far no counter-forensic techniques are published against image forensics based on machine learning, such as camera model identification [16].

On a more general level, also the combination of indicators from several forensic techniques is a kind of dimensionality expansion. Already in their seminal paper, Popescu and Farid [38, p. 146] conjectured:

> *"[. . .] as the suite of detection tools expands we believe that it will become increasingly harder to simultaneously foil each of the detection schemes."*

Nevertheless, it remains the responsibility of researchers in the field to substantiate this conjecture with facts, and measure how hard the counterfeiters' task will become. For this, image forensics has to be understood as a security technique, which is measured by its resistance to attacks.

Future proposals of new forensic techniques should consequently come with a security evaluation. Researchers should try to attack their schemes and argue or demonstrate how effective these attacks can be. This way, the field will develop similar to the field of information hiding, where authors in the 1990s mainly focused

on message integrity and imperceptibility, and rigorous security evaluations became standard in the 2000s. New results on counter-forensics should in no way be understood as "helping the perpetrators", but they will become relevant contributions on their own, just like innovations in cryptanalysis serve as important benchmarks to evaluate new cryptographic systems.

Acknowledgements

Thanks are due to Gang Cao, Thomas Gloe, and Miroslav Goljan for sharing the data of their original research papers.

References

1. Barni, M., Cancelli, G., Esposito, A.: Forensics aided steganalysis of heterogeneous images. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2010, pp. 1690–1693. IEEE (2010)
2. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer-Verlag, Berlin, Heidelberg (2006)
3. Böhme, R.: Advanced Statistical Steganalysis. Springer-Verlag, Berlin, Heidelberg (2010)
4. Böhme, R., Freiling, F., Gloe, T., Kirchner, M.: Multimedia forensics is not computer forensics. In: Z.J. Geradts, K.Y. Franke, C.J. Veenman (eds.) Computational Forensics, Third International Workshop, IWCF 2009, *Lecture Notes in Computer Science*, vol. 5718, pp. 90–103. Springer-Verlag, Berlin, Heidelberg (2009)
5. Cachin, C.: An information-theoretic model for steganography. In: D. Aucsmith (ed.) Information Hiding, Second International Workshop, IH 98, pp. 306–318. Springer-Verlag, Berlin, Heidelberg (1998)
6. Cao, G., Zhao, Y., Ni, R., Tian, H.: Anti-forensics of contrast enhancement in digital images. In: MM&Sec '10, Proceedings of the 12th ACM workshop on Multimedia and Security, pp. 25–34. ACM Press, New York (2010)
7. Cao, G., Zhao, Y., Ni, R., Yu, L., Tian, H.: Forensic detection of median filtering in digital images. In: IEEE International Conference on Multimedia and EXPO, ICME 2010, pp. 89–94. IEEE (2010)
8. Chinese Academy of Sciences, Institute of Automation: Casia tampered image detection evaluation database (2009/2010). URL `http://forensics.idealtest.org`
9. Christlein, V., Riess, C., Angelopoulou, E.: A study on features for the detection of copy-move forgeries. In: F.C. Freiling (ed.) Sicherheit 2010, pp. 105–116. Gesellschaft für Informatik e.V., Bonn (2010)
10. Cox, I.J., Miller, M.L., Bloom, J.A., Fridrich, J., Kalker, T.: Digital Watermarking and Steganography. Morgan Kaufmann (2008)
11. Cromey, D.W.: Avoiding twisted pixels: Ethical guidelines for the appropriate use and manipulation of scientific digital images. Science and Engineering Ethics **16**(4), 639–667 (2010)
12. Fan, Z., Queiroz, R.L.: Identification of bitmap compression history: JPEG detection and quantizer estimation. IEEE Transactions on Image Processing **12**(2), 230–235 (2003)
13. Fridrich, J.: Digital image forensics. IEEE Signal Processing Magazine **26**(2), 26–37 (2009)
14. Fridrich, J.: Steganography in Digital Media: Principles, Algorithms, and Applications. Cambridge University Press, New York, NY (2009)
15. Friedman, G.: The Trustworthy Digital Camera: Restoring Credibility to the Photographic Image. IEEE Transactions on Consumer Electronics **39**(4), 905–910 (1993)

16. Gloe, T., Borowka, K., Winkler, A.: Feature-based camera model identification works in practice: Results of a comprehensive evaluation study. In: S. Katzenbeisser, A.R. Sadeghi (eds.) Information Hiding, 11th International Workshop, IH 2009, *Lecture Notes in Computer Science*, vol. 5806, pp. 262–276. Springer-Verlag, Berlin, Heidelberg (2009)
17. Gloe, T., Kirchner, M., Winkler, A., Böhme, R.: Can we trust digital image forensics? In: MULTIMEDIA '07, Proceedings of the 15th International Conference on Multimedia, pp. 78–86. ACM Press, New York, NY, USA (2007)
18. Goljan, M., Fridrich, J., Chen, M.: Sensor noise camera identification: Countering counter-forensics. In: N.D. Memon, J. Dittmann, A.M. Alattar, E.J. Delp (eds.) Media Forensics and Security II, *Proceedings of SPIE*, vol. 7541, p. 75410S. SPIE, Bellingham, WA (2010)
19. Hsu, Y.F., Chang, S.F.: Detecting image splicing using geometry invariants and camera characteristics consistency. In: IEEE International Conference on Multimedia and EXPO, ICME 2006, pp. 549–552. IEEE (2006)
20. Huang, Y.: Can digital image forgery detection be unevadable? A case study: color filter array interpolation statistical feature recovery. In: S. Li, F. Pereira, H.Y. Shum, A.G. Tescher (eds.) Proceedings of SPIE: Visual Communications and Image Processing, vol. SPIE 5960, p. 59602W (2005)
21. Johnson, M.K., Farid, H.: Exposing digital forgeries in complex lighting environments. IEEE Transactions on Information Forensics and Security **2**(3), 450–461 (2007)
22. Kalker, T.: Considerations on watermarking security. In: IEEE Fourth Workshop on Multimedia Signal Processing, pp. 201–206. IEEE (2001)
23. Kerckhoffs, A.: La cryptographie militaire. Journal des sciences militaires **IX**, 5–38, 161–191 (1883)
24. Kirchner, M., Böhme, R.: Tamper hiding: Defeating image forensics. In: T. Furon, F. Cayre, G. Doërr, P. Bas (eds.) Information Hiding, 9th International Workshop, IH 2007, *Lecture Notes in Computer Science*, vol. 4567, pp. 326–341. Springer-Verlag, Berlin, Heidelberg (2007)
25. Kirchner, M., Böhme, R.: Hiding traces of resampling in digital images. IEEE Transactions on Information Forensics and Security **3**(4), 582–592 (2008)
26. Kirchner, M., Böhme, R.: Synthesis of color filter array pattern in digital images. In: E.J. Delp, J. Dittmann, N.D. Memon, P.W. Wong (eds.) Media Forensics and Security, *Proceedings of SPIE*, vol. 7254, p. 72540K. SPIE, Bellingham, WA (2009)
27. Kirchner, M., Fridrich, J.: On detection of median filtering in digital images. In: N.D. Memon, J. Dittmann, A.M. Alattar, E.J. Delp (eds.) Media Forensics and Security II, Proceedings of SPIE-IS&T Electronic Imaging, vol. SPIE 7541, p. 754110. San Jose, CA, USA (2010)
28. Lai, S., Böhme, R.: Countering counter-forensics: The case of JPEG compression. In: T. Filler, T. Pevný, S. Craver, A. Ker (eds.) Information Hiding, 13th International Conference, IH 2011, vol. 6958, pp. 285–298. Springer-Verlag, Berlin, Heidelberg (2011)
29. Lam, E.Y., Goodman, J.W.: A mathematical analysis of the DCT coefficient distributions for images. IEEE Transactions on Image Processing **9**(10), 1661–1666 (2000)
30. Li, C.T., Chang, C.Y., Li, Y.: On the repudiability of device identification and image integrity verification using sensor pattern noise. In: D. Weerasinghe (ed.) Information Security and Digital Forensics, First International Conference, ISDF 2009, *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, vol. 41, pp. 19–25. Springer-Verlag, Berlin, Heidelberg (2010)
31. Liew, A.W.C., Yan, H.: Blocking artifacts suppression in block-coded images using overcomplete wavelet representation. IEEE Transactions on Circuits and Systems for Video Technology **14**(4), 450–461 (2004)
32. Lukáš, J., Fridrich, J., Goljan, M.: Detecting digital image forgeries using sensor pattern noise. In: A. Said, J.G. Apostolopoulos (eds.) Proceedings of SPIE: Image and Video Communications and Processing 2005, vol. 5685, pp. 249–260 (2005)
33. Lukáš, J., Fridrich, J., Goljan, M.: Digital camera identification from sensor noise. IEEE Transactions on Information Forensics and Security **1**(2), 205–214 (2006)
34. Meerwald, P., Uhl, A.: Additive spread-spectrum watermark detection in demosaicked images. In: MM&Sec '09, Proceedings of the Multimedia and Security Workshop 2009, pp. 25–32 (2009)

35. Ng, T.T., Chang, S.F.: A data set of authentic and spliced image blocks. Tech. Rep. ADVENT 203-2004-3, Department of Electrical Engineering, Columbia University, New York, NY, USA (2004)
36. Parrish, D., Noonan, B.: Image manipulation as research misconduct. Science and Engineering Ethics **15**(2), 161–167 (2008)
37. Pfitzmann, A., Hansen, M.: A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management. `http://dud.inf.tu-dresden.de/Anon_Terminology.shtml` (2010). (Version 0.34)
38. Popescu, A.C., Farid, H.: Statistical tools for digital forensics. In: J. Fridrich (ed.) Information Hiding, 6th International Workshop, IH 2004, *Lecture Notes in Computer Science*, vol. 3200, pp. 128–147. Springer-Verlag, Berlin, Heidelberg (2004)
39. Popescu, A.C., Farid, H.: Exposing digital forgeries in color filter array interpolated images. IEEE Transactions on Signal Processing **53**(10), 3948–3959 (2005)
40. Sallee, P.: Model-based methods for steganography and steganalysis. International Journal of Image and Graphics **5**(1), 167–189 (2005)
41. Stamm, M.C., Liu, K.J.R.: Blind forensics of contrast enhancement in digital images. In: IEEE International Conference on Image Processing, ICIP 2008, pp. 3112–3115. IEEE (2008)
42. Stamm, M.C., Liu, K.J.R.: Wavelet-based image compression anti-forensics. In: IEEE International Conference on Image Processing, ICIP 2010. IEEE (2010)
43. Stamm, M.C., Tjoa, S.K., Lin, W.S., Liu, K.J.R.: Anti-forensics of JPEG compression. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2010, pp. 1694–1697. IEEE (2010)
44. Stamm, M.C., Tjoa, S.K., Lin, W.S., Liu, K.J.R.: Undetectable image tampering through JPEG compression anti-forensics. In: IEEE International Conference on Image Processing, ICIP 2010. IEEE (2010)
45. Steinebach, M., Liu, H., Fan, P., Katzenbeisser, S.: Cell phone camera ballistics: Attacks and countermeasures. In: R. Creutzburg, D. Akopian (eds.) Multimedia on Mobile Devices 2010, *Proceedings of SPIE*, vol. 7542, p. 75420B. SPIE, Bellingham, WA (2010)
46. Steinebach, M., Ouariachi, M.E., Liu, H., Katzenbeisser, S.: On the reliability of cell phone camera fingerprint recognition. In: S. Goel (ed.) Digital Forensics and Cyber Crime, First International ICST Conference, ICDF2C 2009, *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, vol. 31, pp. 69–76. Springer-Verlag, Berlin, Heidelberg (2009)
47. Thévenaz, P., Blu, T., Unser, M.: Image interpolation and resampling. In: Handbook of medical imaging, pp. 393–420. Academic Press, Orlando, FL (2000)
48. Ullerich, C., Westfeld, A.: Weaknesses of MB2. In: Y.Q. Shi, H.J. Kim, S. Katzenbeisser (eds.) Digital Watermarking (Proceedings of IWDW 2007), *Lecture Notes in Computer Science*, vol. 5041, pp. 127–142. Springer-Verlag, Berlin, Heidelberg (2008)
49. Valenzise, G., Nobile, V., Tagliasacchi, M., Tubaro, S.: Countering JPEG anti-forensics. In: IEEE International Conference on Image Processing, ICIP 2011, pp. 1949–1952. IEEE (2011)
50. Valenzise, G., Tagliasacchi, M., Tubaro, S.: The cost of JPEG compression anti-forensics. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2011, pp. 1884–1887. IEEE (2011)
51. Yuan, H.D.: Blind forensics of median filtering in digital images. IEEE Transactions on Information Forensics and Security **6**(4), 1335–1345 (2011)
52. Zhai, G., Zhang, W., Yang, X., Lin, W., Xu, Y.: Efficient image deblocking based on postfiltering in shifted windows. IEEE Transactions on Circuits and Systems for Video Technology **18**(1), 122–126 (2008)