Can We Trust Digital Image Forensics?

Thomas Gloe
Technische Universität Dresden
Institute for System Architecture
01062 Dresden, Germany
thomas.gloe@inf.tu-dresden.de

Matthias Kirchner
Technische Universität Dresden
Institute for System Architecture
01062 Dresden, Germany
matthias.kirchner@inf.tu-dresden.de

Antje Winkler
Technische Universität Dresden
Institute for System Architecture
01062 Dresden, Germany
antje.winkler@inf.tu-dresden.de

Rainer Böhme
Technische Universität Dresden
Institute for System Architecture
01062 Dresden, Germany
rainer.boehme@inf.tu-dresden.de

ABSTRACT

Compared to the prominent role digital images play in nowadays multimedia society, research in the field of image authenticity is still in its infancy. Only recently, research on digital image forensics has gained attention by addressing tamper detection and image source identification. However, most publications in this emerging field still lack rigorous discussions of robustness against strategic counterfeiters, who anticipate the existence of forensic techniques. As a result, the question of trustworthiness of digital image forensics arises. This work will take a closer look at two state-of-the-art forensic methods and proposes two counter-techniques; one to perform resampling operations undetectably and another one to forge traces of image origin. Implications for future image forensic systems will be discussed.

Categories and Subject Descriptors

I.4 [**Image Processing**]: Miscellaneous

General Terms

Algorithms, Security

Keywords

Tamper Hiding, Tamper Detection, Image Source Identification, Digital Image Forensics

1. INTRODUCTION

Back in analog times, a photograph was generally perceived as a "piece of truth". With digital image processing replacing its analog counterpart, critics have expressed the concern that it has never been so easy to manipulate images. The advent of low-cost digital imaging devices as well as powerful and sophisticated editing software makes it no longer necessary to obtain specialist skills to alter an image's tenor. Thus, questions regarding image authenticity are of growing relevance, especially in contexts where nowadays' multimedia society bases important decisions on them. Lately discovered forgeries in newspapers and scientific journals are only the tip of the iceberg. Particular attention has to be drawn to courtroom applications, in which the authenticity of photographs as pieces of evidence deserves utmost importance.

Recently, methods subsumed to the concept of *digital image forensics* have been proposed to address these issues. The area of digital image forensics can be broadly divided into two branches [14]. The first field of application is to determine whether a specific digital image has undergone malicious post-processing or tampering. Forensic algorithms of this type are designed to unveil either characteristic traces of image processing operations, or to verify the integrity of particular features introduced in a typical image acquisition process. The second problem linked to digital image forensics is image source identification, which is obviously based on specific characteristics of the image acquisition device or technology. As forensic algorithms basically rely on particular statistical features, which can be understood as a "natural" and inherent watermark, digital image forensics does not require any prior knowledge of the original image.

Since in general, existing methods are deemed quite reliable in laboratory tests, one might be tempted to apply them in practice as well. However, little is known about the robustness of forensic algorithms. This aspect plays only a marginal role in the existing body of literature. As a consequence, it is reasonable to question the trustworthiness of digital image forensics — in particular with regard to a farsighted counterfeiter who is aware of forensic tools.

To draw attention to this disproportion, this paper focuses on two specific forensic methods — a resampling detector proposed by Popescu and Farid [16] and an approach to digital camera identification by Lukáš, Fridrich and Goljan [11] — and develops ways to deceive these methods. Forensic methods might benefit from research on countermeasures in a similar way as reasoning about attacks in multimedia security in general is useful to improve security. In this sense, attacks on image forensic algorithms can be understood as schemes to systematically mislead the detection methods.

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany.
Copyright 2007 ACM 978-1-59593-701-8/07/0009 ...$5.00.

In general, such attacks can be assigned to one of the following three objectives, namely:

1. the camouflage of malicious post-processing or tampering of an image,
2. the suppression of correct image origin identification,
3. and furthermore, the forgery of image origin.

The paper is organized as follows: Section 2 briefly reviews commonly known image forensic algorithms in general and describes in more detail the two specific forensic algorithms under investigation. Section 3 points out possible weaknesses of digital image forensics and explains two countertechniques. The first one allows to perform resampling operations undetectably while the second technique aims at misleading image source identification methods. The effectiveness of both techniques is validated with experimental results. Finally, Section 4 summarizes and provides an outlook for future work.

2. DIGITAL IMAGE FORENSICS

Digital image forensic techniques exploit either traces of image processing algorithms or characteristics introduced during the image acquisition process. The former are applicable without knowledge about the used digitization device. For example, a method proposed by Popescu and Farid reveals dependencies introduced during resizing or rotation of images [16]. Other methods use, for example, statistics of JPEG coefficients to detect recompression [10], or analyze phase congruency to detect image splicing [1].

To illustrate some characteristics typically introduced during image acquisition, Figure 1 shows a simplified image processing pipeline of a digital camera. The main components are the lens, the sensor with a color filter array (CFA) and the signal processing unit. The CFA is needed for color images as typical sensors are only sensitive to the intensity of incoming light. A true color RGB-image is obtained from interpolating intensity values of pixels in a close neighborhood.

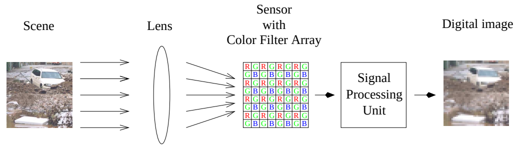

Figure 1: Image acquisition process in a digital camera.

The captured image data is further processed in the signal processing unit and afterwards stored in a data storage unit. Other digital image input devices, such as digital camcorders or digital flatbed scanners, use similar image processing pipelines and thus introduce similar statistical patterns in the image data.

Forensic algorithms may exploit specific characteristics of image statistics, which were introduced by components of the image processing pipeline. Starting with the lens, chromatic aberration [6] and radial distortions [2] are adequate features. Furthermore, defect sensor elements [4], sensor noise [11] and dependencies between adjacent pixels due to color interpolation [17] form typical ingredients for forensic methods. However, it is also possible to consider the whole image acquisition process as a black box and analyze the camera response function [9] or macroscopic features of acquired images [7].

Among the number of image forensic techniques proposed in the literature, we have selected two very important and useful algorithms, which we will briefly describe in the following section before demonstrating, in the remainder of the paper, how to deceive these techniques.

2.1 Detecting Traces of Resampling

A typical setup for creating convincing digital image forgeries often involves resizing and rotation of images or parts thereof. The transformed image $\mathbf{y}$ then results from resampling the original image $\mathbf{x}$ to a new image lattice. Assuming the commonly used affine transformation, each pixel's position in the transformed image is computed from the following transformation relation:

$$\begin{bmatrix} i_{\mathbf{y}} \\ j_{\mathbf{y}} \end{bmatrix} = \mathbf{A} \cdot \begin{bmatrix} i_{\mathbf{x}} \\ j_{\mathbf{x}} \end{bmatrix} \ . \tag{1}$$

$\mathbf{A}$ is the 2×2 transformation matrix and indices $i_{\mathbf{x}}, j_{\mathbf{x}}$ refer to source positions as opposed to $i_{\mathbf{y}}, j_{\mathbf{y}}$, which index the resampled image. In the general case, interpolation is required to match the real-valued transformed positions with the integer grid of the target image and thus enable smooth and visually appealing image transformations [20]. While these transformations might be imperceptible at first glance, a forensic investigator can benefit from the fact that interpolation introduces linear dependencies between groups of adjacent pixels. The state-of-the-art resampling detector, as proposed by Popescu and Farid, is designed to unveil these artifacts in transformed bitmap images [16]. Therefore, their detection scheme uses a linear model to approximate a pixel's intensity as the weighted sum of pixels in its close neighborhood (window of size $(2N+1) \times (2N+1)$, with N integer) and an independent residual $\boldsymbol{\epsilon}$,

$$y_{i,j} = f(\boldsymbol{\alpha}, \mathbf{y}) + \epsilon_{i,j} = \sum_{(k,l) \in \{-N,\ldots,N\}^2} \alpha_{k,l} \cdot y_{i+k,j+l} \ + \ \epsilon_{i,j} \ . \tag{2}$$

Scalar weights $\boldsymbol{\alpha}$ correspond to the specific form of the correlation introduced to the image. As the interpolation is applied to an equidistant sampling lattice, correlation appears in a systematic and periodic manner.

However, in a practical setting it is neither known which pixels are genuinely linear combinations of their neighbors, nor in which way these pixels correlate with their neighbors. Popescu and Farid propose to use the *expectation maximization* (EM) algorithm [3], an iterative two-stage procedure, to simultaneously estimate both each pixel's probability for being a linear combination of its neighbors and the unknown weights $\boldsymbol{\alpha}$. The so-obtained probabilities are referred to as *p-map* and exhibit a conspicuous periodical pattern after previous re-sampling operations. When transformed to the frequency domain (using a discrete Fourier transformation), this pattern causes distinct peaks in the spectrum. The location of the peaks is typical for the specific resampling parameters. To enhance the visibility of the characteristic peaks, Popescu and Farid propose to apply a contrast function. The contrast function is composed of a radial weighting window, which attenuates very low frequencies, and a gamma correction step.

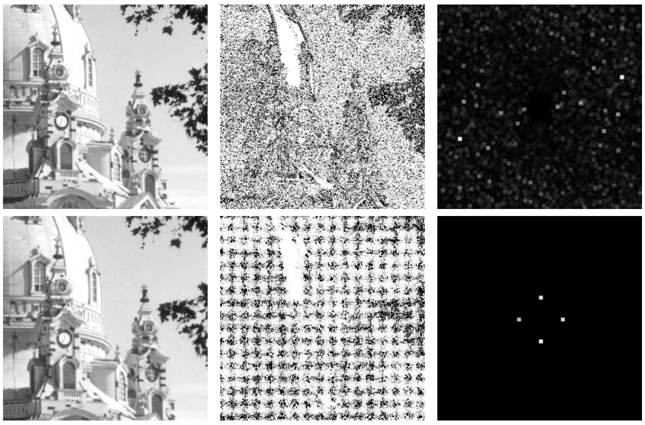

Figure 2: Resampling detection results for an original image (top row) and a 5 % upsampled version (bottom row). Note that only the transformed image's *p*-map shows a clear periodic pattern (middle column), which results in characteristic peaks in the map's frequency spectrum (right column).

Figure 2 shows the detection results for a typical grayscale image. The detector was applied to both the original image (top row) and a processed version, which had been scaled up with linear interpolation to 105 % of the original (bottom row). The corresponding *p*-maps are displayed in the middle column. While the *p*-map of the original image is rather chaotic, the periodical pattern described above is clearly visible in case of the transformed image. Note the characteristic spectral peaks, which only appear in the manipulated image's *p*-map (right column).

Automatic Detection of Resampling

To automatically identify forgeries in a batch process, it is necessary to quantify the periodic artifacts in a given image. Therefore, Popescu and Farid suggest to measure the similarities between the estimated *p*-map and a set of synthetically generated periodic patterns. They found empirically that a synthetic map $\mathbf{s}^{(\mathbf{A})}$ for a specific transformation $\mathbf{A}$ can be computed as the distance between each point in the resampled lattice and the closest point in the original lattice,

$$s_{i,j}^{(\mathbf{A})} = \left\| \mathbf{A} \cdot \begin{bmatrix} i \\ j \end{bmatrix} - \left\lfloor \mathbf{A} \cdot \begin{bmatrix} i \\ j \end{bmatrix} + \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix} \right\rfloor \right\| . \qquad (3)$$

Once a (sufficiently dimensioned) set $\mathcal{A}$ of candidate transformations $\mathbf{A}$ is defined, an automatic detector uses the maximum pairwise similarity between an estimated *p*-map and all elements of $\mathcal{A}$ as decision criterion δ_{RD} ('RD' for resampling detection),

$$\delta_{\mathrm{RD}} = \max_{\mathbf{A} \in \mathcal{A}} \sum_{i,j} \left| \mathrm{C}(\mathrm{DFT}(\mathbf{p})) \right| \cdot \left| \mathrm{DFT}\left(\mathbf{s}^{(\mathbf{A})} \right) \right| . \qquad (4)$$

Function C is the contrast function and DFT applies a 2D discrete Fourier transformation. If δ_{RD} exceeds a predefined threshold $\delta_{\mathrm{RD}}^{(\mathrm{T})}$, the corresponding image is flagged as resampled.

As demonstrated in the original publication [16], this detection method can be considered as a very reliable and powerful tool to unveil a great variety of geometric image transformations. Robustness against several image processing operations has already been proven and can be confirmed by us.

2.2 Identification of Digital Camera Image Origin

Lukáš, Fridrich and Goljan [11] first proposed to use sensor noise for digital camera identification. Sensor noise is inherently present in each image captured with a digital camera. It consists of two main components: *temporal* and *spatial* noise [19]. However, temporal noise is not suitable for identification, as it is stochastically independent for pixels within one image as well as between images.

Contrary to temporal noise, spatial noise is relatively stable between images and can therefore be used for camera identification. All images acquired with the same image input device contain a similar spatial noise pattern, which Lukáš et al. assume to be unique for each sensor. Spatial noise consists of three components: Photo Response Non-Uniformity Noise (PRNU), Fixed Pattern Noise (FPN) and other irregularities resulting from local disturbances on the optical elements (scratches, dust, etc.).

To estimate the spatial noise for a device under investigation, Lukáš et al. propose to use the residual of a wavelet denoising filter [13]. A camera-specific fingerprint called *reference noise pattern* can be obtained by averaging the estimated noise of about 300 images from the same device. Averaging is necessary to separate temporal from spatial noise.

The similarity of a specific digital image's noise pattern $\hat{\mathbf{n}}$ and a reference noise pattern $\mathbf{r}_c$ is measured in terms of the correlation coefficient $\rho(\mathbf{r}_c, \hat{\mathbf{n}})$,

$$\rho(\mathbf{r}_c, \hat{\mathbf{n}}) = \frac{\sum_{i,j} \left((\hat{\mathbf{n}} - \mathrm{E}[\hat{\mathbf{n}}]) \cdot (\mathbf{r}_c - \mathrm{E}[\mathbf{r}_c]) \right)}{\| \hat{\mathbf{n}} - \mathrm{E}[\hat{\mathbf{n}}] \| \| \mathbf{r}_c - \mathrm{E}[\mathbf{r}_c] \|} . \qquad (5)$$

The correlation coefficient is determined for all candidate cameras $c \in \mathcal{C}$. The image under investigation is assigned to a camera according to a decision criterion δ_{CI} ('CI' for camera identification),

$$\delta_{\mathrm{CI}} = \max_{c \in \mathcal{C}} \rho(\mathbf{r}_c, \hat{\mathbf{n}}) . \qquad (6)$$

If δ_{CI} exceeds a specified threshold $\delta_{\mathrm{CI}}^{(\mathrm{T})}$, the corresponding image is assigned to the camera $\arg \delta_{\mathrm{CI}}$. Note that thresholds are determined empirically and that there is no consensus yet on its optimal level for large-scale applications.

Figure 3 shows example correlation coefficients for noise patterns of images captured with several input devices and the reference pattern of one Canon PowerShot S70 digital camera. As can be seen from the graph, it is possible to correctly identify the image origin for all photographs obtained with the Canon S70 camera in this defined set of input devices.

Lukáš et al. have shown that a correct identification is still possible after resizing an image, after suppression of image noise with the wavelet denoising filter, and after applying JPEG compression. However, this method is sensitive to geometric transformations such as cropping or rotation.

3. ATTACKS ON DIGITAL IMAGE FORENSICS

The previous section provided an overview of the powerful possibilities that digital image forensics offers to the

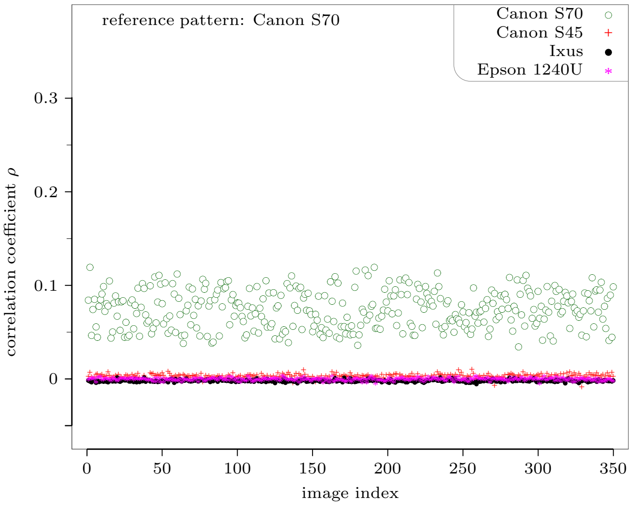

Figure 3: Identification results for digital camera Canon PowerShot S70.

field of image authentication. The presented techniques are currently widely accepted and probably already applied in practice. Forensic extensions to popular commercial image editing software are already announced.[1] However, users of such tools should keep in mind that results from forensic investigations can serve as indications only. Therefore, questions regarding the conclusiveness of image forensics are all the more of high importance. Unfortunately, the available literature is primarily focused on the detectors' sensitivity and rarely considers strategic attacks or countermeasures.

Corresponding to the protection goals of image authentication schemes in general and digital image forensics in particular, an attacker might attempt to cover up manipulative image processing or to prevent the correct identification of the image source. To achieve these goals, an attacker might choose either targeted or universal attacks. A *targeted attack* exploits characteristics and vulnerabilities of a particular forensic technique, which the developer of the attack usually knows. Nevertheless, it is possible that other forensic methods are capable of detecting traces of previous manipulations. To mislead even unknown forensic tools, *universal attacks* aim at providing correct and plausible image statistics. This obviously is the more difficult task, as correct image statistics imply the compliance with stochastic image models, which are not fully understood [18]. Hence, attackers can never be sure whether their forgery is free of detectable artifacts. Since the application of some typical image processing methods is commonly accepted, these methods may serve as means to obtain plausible image statistics while at the same time overriding subtle statistical traces of prior manipulation.

3.1 Universal Attacks in Terms of Commonly Used Image Processing Algorithms

Today's image processing toolboxes offer a great variety of commonly used image filtering and compression algorithms. Since the application of, e.g., noise reduction filters or JPEG compression can be regarded as plausible steps, a robust image forensic method should withstand such additional influences. Both Popescu and Farid [16] and Lukáš et al. [11] reported robustness against several point operations, such as gamma correction or adding noise. Additionally, the latter method has shown good results even for JPEG compressed and resized images. As both presented techniques rely on specific spatial features, the application of neighborhood operations may interfere with detection reliability. For example, the basic median filter, a non-linear and signal adaptive smoothing operator, is likely to impede the correct detection of previous image transformations [8] as well as source identification. This is plausible for the former method because the detection of resampling is directed at finding local linear pixel correlations. It is evident for digital camera identification as well, since this technique relies on fixed camera specific spatial noise patterns.

Moreover, the window size of the applied median filter plays a crucial role for the two forensic algorithms. Larger window sizes introduce a higher degree of non-linearity and thus suppress correct detection more effectively. At the same time, a large filter window goes along with considerable degradation in image quality. A similar decrease of detection accuracy at the expense of low image quality can be observed for several other filter types, such as mean, binomial and Gaussian filters.

Although a counterfeiter may use these or similar techniques to simulate plausible image statistics, the necessary tradeoff between undetectability and actual visual impact highlights that naive post processing approaches are suboptimal for this purpose. Therefore, a farsighted counterfeiter would rather choose a targeted attack on a specific detection method. In the following subsections, an approach to undetectable resampling and an attack against correct identification of digital cameras will be presented.

3.2 An Approach to Undetectable Resampling

The resampling detector proposed by Popescu and Farid relies on finding systematic and periodic dependencies between pixels in a close neighborhood. This periodicity is due to the equidistant sampling lattice. During the mapping of discrete lattice positions from source to destination image, the relative position of source and target pixel is repeated over the entire plane. To break this equidistance and allow for undetectable resampling, we introduce geometric distortions, which are known from watermarking attacks as well (see [15], though in a slightly different variant). Here we superimpose a random disturbance vector $\mathbf{e}$ to each individual pixel's position as calculated from the transformation relation,

$$\begin{bmatrix} i_{\tilde{\mathbf{y}}} \\ j_{\tilde{\mathbf{y}}} \end{bmatrix} = \mathbf{A} \cdot \begin{bmatrix} i_{\mathbf{x}} \\ j_{\mathbf{x}} \end{bmatrix} + \begin{bmatrix} e_{1,i,j} \\ e_{2,i,j} \end{bmatrix} \quad \text{where } e \sim \mathcal{N}(0,\sigma) \text{ i.i.d.} \qquad (7)$$

Parameter σ controls the degree of distortion introduced to the image lattice. Unfortunately, naive geometric distortion is likely to generate visible artifacts, such as jitter, in the resulting image. This is particularly visible at straight lines and edges. To reduce such undesirable side effects, the strength of distortion is adaptively modulated by the local image content. A slight modification of equation (7) employs two edge detectors to control the modulation,

$$\begin{bmatrix} i_{\tilde{\mathbf{y}}} \\ j_{\tilde{\mathbf{y}}} \end{bmatrix} = \mathbf{A} \cdot \begin{bmatrix} i_{\mathbf{x}} \\ j_{\mathbf{x}} \end{bmatrix} + \begin{bmatrix} e_{1,i,j} \cdot (1 - {}^{1}\!/_{255} \cdot \text{sobelH}(\mathbf{y}, i_{\mathbf{y}}, j_{\mathbf{y}})) \\ e_{2,i,j} \cdot (1 - {}^{1}\!/_{255} \cdot \text{sobelV}(\mathbf{y}, i_{\mathbf{y}}, j_{\mathbf{y}})) \end{bmatrix} \qquad (8)$$

[1] Adobe tackles photo forgeries: `http://www.wired.com/gadgets/digitalcameras/news/2007/03/72883`

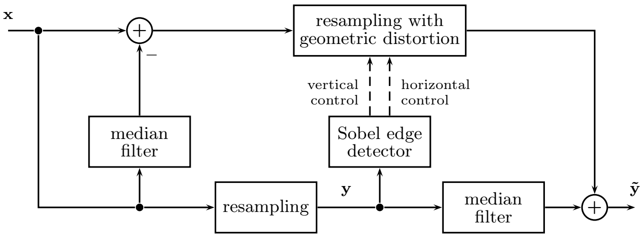

Figure 4: Block diagram for the dual path approach to undetectable resampling.

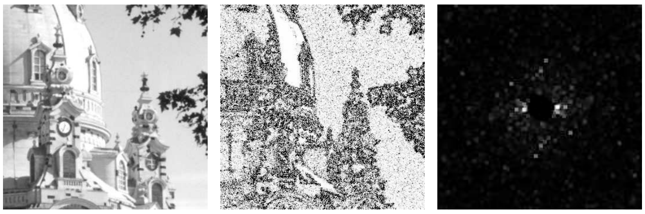

Figure 5: Detection results for 5 % upsampling using the dual path approach. 7 × 7 median filter, $\sigma = 0.3$. The estimated p-map shows no periodic pattern. Characteristic peaks in its spectrum have vanished.

Functions sobelH and sobelV return the value of a linear Sobel filter for horizontal and vertical edge detection, respectively. This construction ensures that pronounced edges in vertical direction cause less distortion in horizontal direction and vice versa. Note that at this stage of our attack the same original image has to be resampled twice, the first time without distortion to apply the Sobel filters and subsequently with distortion to generate the final image $\tilde{\mathbf{y}}$ (see [8] for more details on this attack).

A further improvement of our attack is motivated by the observation that resampling cannot be detected in saturated or very homogenous image regions. Equal or virtually identical pixel intensities suppress the formation of periodic artifacts. Thus, it is primarily the high frequency image component, which is of interest from the detector's point of view. Therefore, our best performing attack, which can be seen as a dual path approach (Figure 4), aims at avoiding traces of resampling in this specific image component. Using a simplified image model, the low frequency component of the output image is obtained in a first step by applying a median filter directly to the resampled image $\mathbf{y}$. Second, a high frequency component is extracted from the source image $\mathbf{x}$ by subtracting the median filtered version of this image. This component is then resampled with geometric distortion and edge modulation (see Equation (8)). The edge information is extracted from the resampled image prior to the median filter. The final image $\tilde{\mathbf{y}}$ can be obtained by adding up both components.

Figure 5 shows the detection results for a 5 % upsampled grayscale image, which has been composed according to the attack described above. Note that the estimated p-map shows no conspicuous periodic artifacts and, as a result, no characteristic peaks appear in its spectrum.

Experimental Results

For a quantitative evaluation of our attack against resampling detection, a test database of 200 never compressed 8-bit grayscale images was built. All images were taken with a Canon PowerShot S70 digital camera at full resolution (3112×2328). From each photograph we cropped a region of 852×852 pixels and finally downsampled this part by factor two to avoid possible interferences from periodic patterns which might stem from a color filter array (CFA) interpolation inside the camera [17]. In order to simulate typical image manipulations, a subset of 100 images was upscaled and downscaled by varying amounts using linear interpolation. For our experiments, a set of altogether 384 synthetic maps was created, 256 for upsampling in the range of 1 % to 100 % and 128 for downsampling in the range of 1 % to 50 %, in each case in equidistant steps of 0.4 percentage points. The detector's threshold $\delta_{\mathrm{RD}}^{(T)}$ was determined empirically for a defined false acceptance rate (FAR) by applying the detector to all 200 original images in the database. Our performance measures are detection rates, i.e., the fraction of correctly detected manipulations, for FAR < 1 % and FAR < 50 %, respectively. Since the application of geometric distortion inevitably affects the inherent image structure, we furthermore measure the amount of image degradation in terms of peak signal to noise ratio (PSNR).

Figure 6 illustrates the general efficiency of geometric distortion as a tool for undetectable resampling. The detection rates for images upsampled to both an unaltered, i.e., equidistant, and a distorted sampling lattice ($\sigma = 0.4$) are shown on the left. While upsampling is perfectly detectable without attack, detectability drops substantially if we distort the image lattice. As can be seen from the graphs, using edge modulation is a reasonable extension to the general approach. While the detection rates decrease even more in case of using edge modulation, the PSNR values indicate a considerable improvement in average image quality of up to 6 dB (right).

Figure 7 depicts the results for the modified attack (Figure 4), which suppresses traces of resampling in the transformed image's high frequency component. The top row reports detection rates for upsampling (left) and downsampling[2] (right) with distortion strength $\sigma = 0.4$. Median filters of size 5×5 and 7×7 have been used to separate the frequency components. Average PSNR values for $\sigma = 0.4$ are shown in the bottom row. Although both 5×5 and 7×7 median filters lead to similar detection rates, the former might be preferred with regard to average image quality. Generally, image quality metrics indicate only marginal loss for the high frequency adjusted attack when compared to simple geometric distortions (with edge modulation). At the same time, the former approach achieves better undetectablity and thus can be deemed as advantageous.

The very low detection rates as presented in Figure 7 demonstrate how successful resampling operations can be concealed with the proposed method. At a practically relevant false acceptance rate of < 1 %, only about 11 % of all image transformations were correctly flagged as resampled (5×5 median filter). Note that the few successful detections were concentrated within just a couple of original images.

[2] Note that unlike upsampling, plain downsampling is not detected perfectly. This is plausible since downsampling causes information loss.

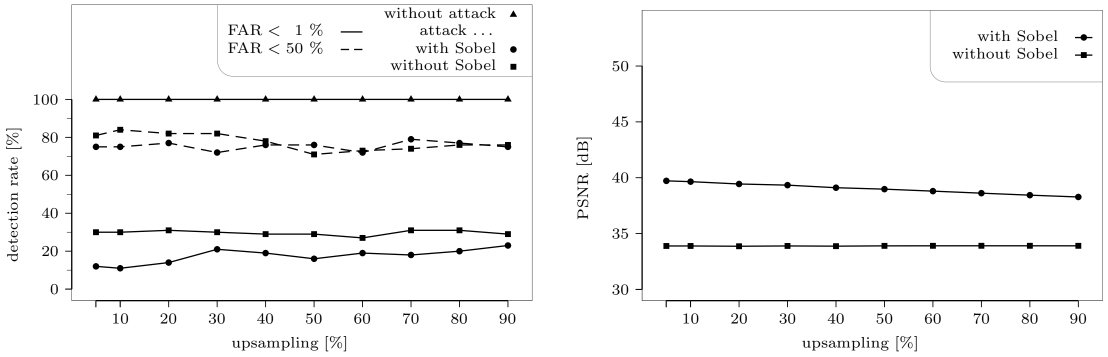

Figure 6: Evaluation of geometric distortion ($\sigma = 0.4$) as tool for undetectable resampling. Detection rates (left) and average image quality after attack (right). While upsampling in the absence of any attack is perfectly detected, geometric distortion causes a substantial drop of detection rates. Edge modulation yields higher average image qualities.

This suggests that image-specific factors may determine the effectiveness of the attack.

3.3 Manipulating Image Source Identification

As the identification method by Lukáš et al. is based on the extraction of a spatial noise pattern, an obvious approach to attack this method relies on suppressing the image's spatial noise. First investigations have shown that it is not sufficient to apply the wavelet denoising filter [13] to defeat the correct identification of image origin [11]. The filter also degrades the visual image quality. Another well-known method to minimize spatial noise of an image is flatfielding [12], which is typically used in astronomy or in flatbed scanners to enhance the final image quality.

The process of flatfielding estimates the two main components of a digital camera's spatial noise, Fixed Pattern Noise (FPN) and Photo Response Non-Uniformity (PRNU), independently. FPN is a signal independent additive noise source. It can be estimated in terms of dark frame $\mathbf{d}$ by averaging K images $\mathbf{x}_{\text{dark}}$ captured in a completely dark environment[3],

$$\mathbf{d} = \frac{1}{K} \sum_{K} \mathbf{x}_{\text{dark}} \, . \tag{9}$$

In contrast to FPN, PRNU is a signal dependent multiplicative noise source. To estimate PRNU, L images of a homogeneously illuminated scene $\mathbf{x}_{\text{light}}$ are required. For these images, the realization of FPN is corrected by subtracting the dark frame $\mathbf{d}$. Subsequently, the corrected images are averaged to determine the flatfield frame $\mathbf{f}$,

$$\mathbf{f} = \frac{1}{L} \sum_{L} (\mathbf{x}_{\text{light}} - \mathbf{d}). \tag{10}$$

Note that averaging is required for both dark and flatfield frame generation in order to reduce the influence of temporal noise.

Impeding Correct Identification of Image Origin

Equipped with the possibility to extract an estimate of FPN and PRNU of a digital camera, an attacker can try to suppress a correct source identification [12]. Therefore, an image $\mathbf{x}$ taken with a specific camera has to be corrected by means of the corresponding frames $(\mathbf{d}, \mathbf{f})$,

$$\tilde{\mathbf{x}} = \frac{\mathbf{x} - \mathbf{d}}{\mathbf{f}} \, . \tag{11}$$

In practice, features derived from local disturbances (like dust and scratches) might also be taken into account [5]. However, in case of digital camera identification local disturbances seem to be negligible.

Because it is difficult to generate an uniformly illuminated scene inside a digital camera, flatfielding is typically not applied automatically. Furthermore, the application of *perfect* flatfielding on a large number of images is difficult, since all parameters (e.g., exposure time) used for estimating FPN and PRNU must match the parameters of the target image. Thus, dark and flatfield frames for all possible parameter combinations would be required. For our investigations, however, we work in a simplified scenario, in which dark and flatfield frames were only created for one fixed parameter set[4], while different parameters were used for the acquisition of test images.

Many digital cameras allow to store the acquired images as RAW, TIFF or JPEG compressed images. While RAW image data offers a maximum of control over the final appearance, some cameras provide access only to already color interpolated image data in formats like TIFF or JPEG. However, the application of flatfielding on color interpolated image data is less accurate due to the smoothing of spatial noise during the CFA interpolation.

In order to analyze the influence of noise reduction on the correct identification of image origin, we applied flatfielding to both RAW and TIFF images. The full resolution images of our test data base originated from three digital cameras (Canon PowerShot S45, Canon PowerShot S70 and Canon Ixus IIs) and one Epson Perfection 1240U flatbed scanner. The reference pattern for each image input device was computed from 300 images. Both dark and flatfield frames were calculated from 20 images each.

Figure 8 shows the results for RAW images and the Canon S70 reference pattern. Note that we use two separate sets of images to extract the reference pattern and to conduct

[3] The completely dark environment can be emulated by covering the lens.

[4] The exposure time was set to 1/60 sec, the shutter speed to 5.0 and the ISO speed to 50.

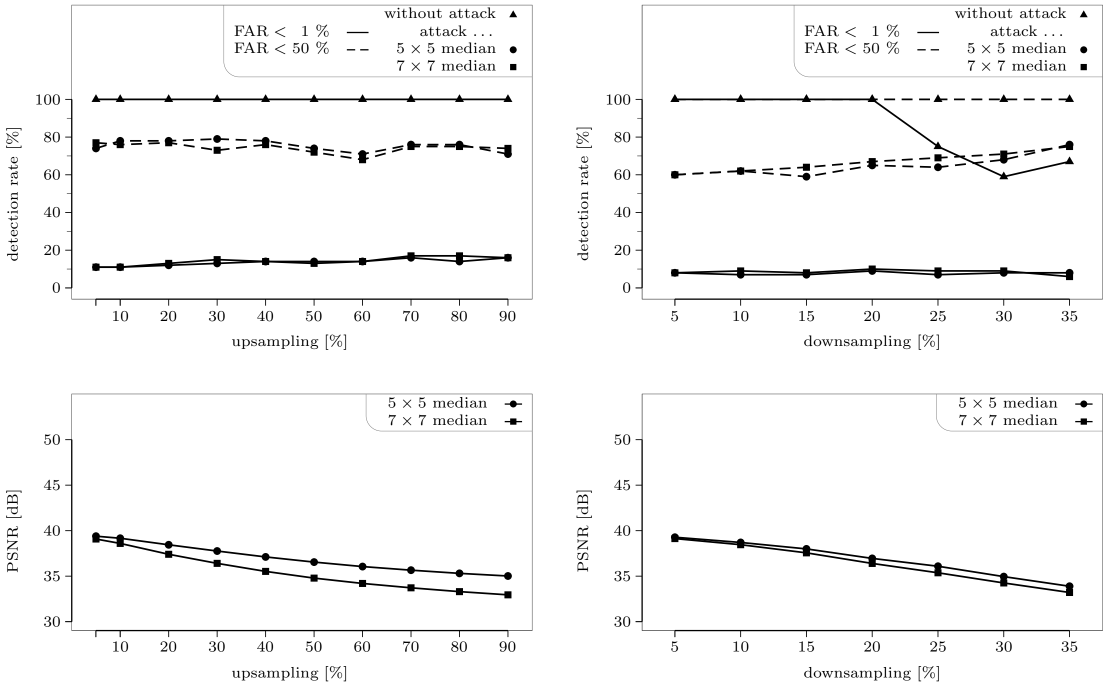

Figure 7: Evaluation of the dual path approach for upsampling (left column) and downsampling (right column). Detection rates (top row) and average image qualities after attack (bottom row) for $\sigma = 0.4$. The size of the applied median filter has minor influence on detection accuracy, whereas a smaller window size preserves better image quality.

the identification tests. As can be seen from the graph, the calculated correlation coefficients for flatfielded images decreased considerably. This means that the characteristic spatial noise could be suppressed to a very low level for all images of our test set, though the correct identification of image origin was successfully prevented only for a subset of images. Not depicted for the sake of brevity are slightly inferior results we achieved for flatfielding of color interpolated image data (also using interpolated data to estimate **d** and **f**).

Since flatfielding is actually used as an illumination correction method, it causes differences between original and processed image. Consequently, the PSNR measure for image quality shows discrepancies of about 30 dB. These values, however, are misleading as flatfielding is known to subjectively enhance (rather than degrade) the visual image quality, as opposed to unwanted side-effects of other image processing operations.

Forging Digital Image Origin

While naive flatfielding merely suppresses spatial noise and therefore allows to disturb the correct identification of image origin, an attacker may also aim at forging the image source. Once the camera's specific noise pattern has been removed by flatfielding, a different pattern can be added to the image by inverse flatfielding,

$$\tilde{\mathbf{y}} = \tilde{\mathbf{x}} \cdot \mathbf{f}_{\text{forge}} + \mathbf{d}_{\text{forge}} \ . \tag{12}$$

The frames $(\mathbf{d}_{\text{forge}}, \mathbf{f}_{\text{forge}})$ correspond to the feigned digital camera, i.e., the application of inverse flatfielding requires access to the flatfielding frames of the corresponding digital camera. However, flatfielding frames might also be approximated from sets of images available on the Internet, e.g., by drawing from large public image data bases like `flickr.com`. To obtain dark and flatfield frames, an attacker could match several homogenous image parts. Note that forging the origin of digital images ideally requires cameras with the same physical resolution. In case of pretending a camera with a smaller resolution, images have to be cropped. Contrary,

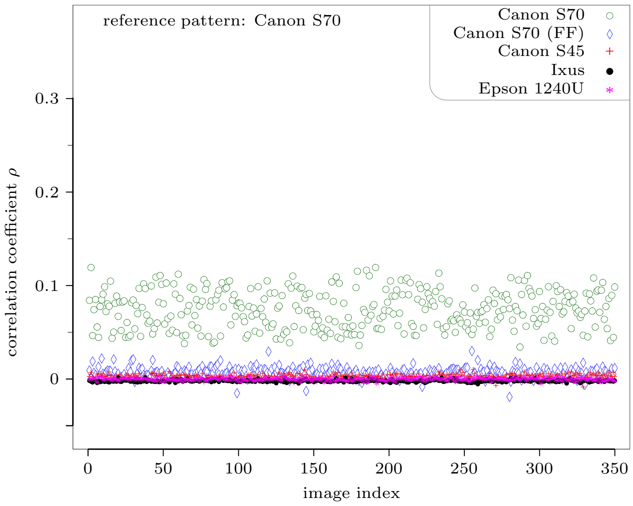

Figure 8: Influence of flatfielding (FF) on digital camera identification (raw image data). Flatfielding suppresses the camera-specific spatial noise pattern.

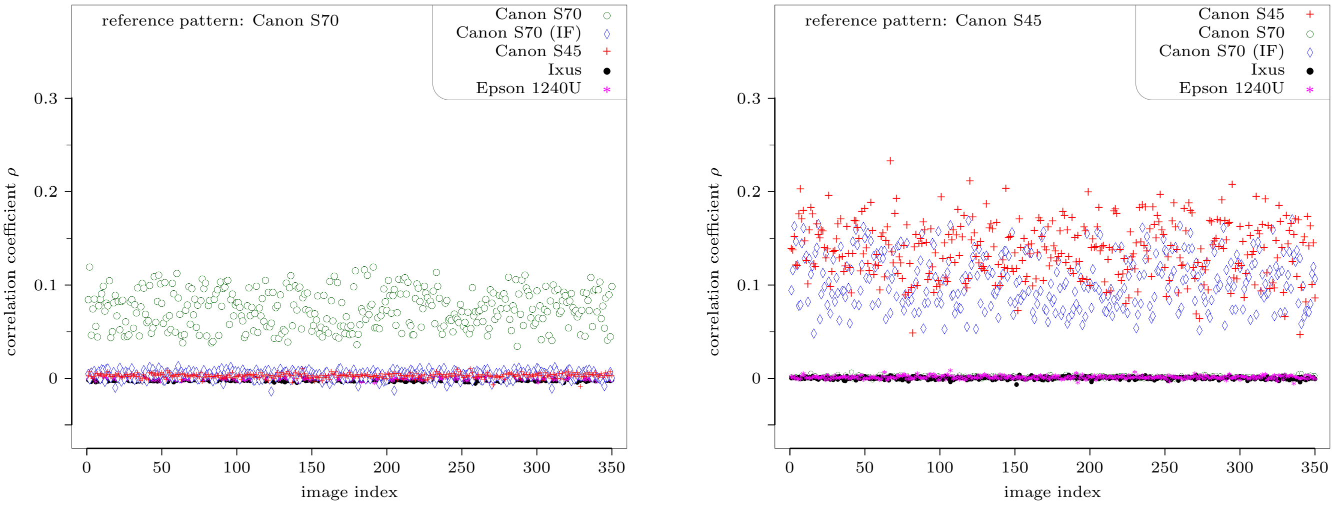

Figure 9: Influence of inverse flatfielding (IF) on digital camera identification (raw image data) for reference patterns corresponding to the actual image origin (left) and the forged image origin (right). Inverse flatfielded images are assigned to forged image origin rather than to the original one.

forging a higher resolution camera results in incompatible image sizes since the resulting images are of a smaller dimension than images typically acquired with this camera.

Figure 9 depicts the results for inverse flatfielding. Using the same test set as described above, the origin of images from the Canon PowerShot S70 was forged to appear as images from the Canon PowerShot S45. The left graph shows the correlation coefficients between the Canon S70 reference pattern, and both the original Canon S70 images and the inverse flatfielded images. The results are similar to those after naive flatfielding, i.e., it is impossible to detect the genuine image source reliably. As the attack aimed at feigning a Canon S45 origin, the right graph shows the correlation coefficients for the Canon S45 reference pattern. The results demonstrate that it is possible to forge image origin, since the correlation coefficients for the forged images are comparable to those of true Canon S45 images.

Analog to the case of naive flatfielding, inverse flatfielding works also for interpolated data. However, as this simplified approach to flatfielding covered only one fixed parameter set to generate dark and flatfield frames, the attack was not able to completely suppress all traces of the original camera's reference pattern. Thus, a more accurate flatfielding has to be applied when using this forgery method in practice.

4. CONCLUDING REMARKS

This paper has taken a critical view on the reliability of forensic techniques as tools to generate evidence of authenticity for digital images. Corresponding to the objective of image authentication schemes, three possible goals of attacks were pointed out. We have discussed an approach for image manipulations undetectable to the resampling detector of Popescu and Farid [16]. Moreover, we have demonstrated how to suppress a reliable image source identification with Lukáš et al.'s method [11], and we have shown a way to falsely frame an 'innocent' camera in the source identification scheme.

More precisely, the following techniques stand behind these results: To avoid typical traces of geometric image transformations in terms of systematic dependencies between adjacent pixels, geometric distortions during the resampling process were introduced. We found that a dual path approach designed to suppress traces of resampling in the transformed image's high frequency component only is particularly promising. Moreover, flatfielding can be used as a method to reduce spatial noise, which is the key to camera identification. While naive flatfielding allows to impede the correct identification of the image source, inverse flatfielding can be used to forge an image origin. Apart from their ability of misleading the respective forensic tools, the proposed attacks do not introduce visual artifacts.

It is well-known from the security community that development of attacks fruitfully helps to improve the security of next generation algorithms. As we designed our attacks in terms of targeted attacks, it is obvious that our results are limited to the specific detection methods under investigation. Thus, it is still possible that further improved detection methods will be able to cope with our attacks. However, such improved detection schemes can be subject to even better targeted attacks by farsighted counterfeiters, what might trigger a cat-and-mouse game between forensic and counter-forensic techniques.

Furthermore, in practice, forensic evidence will often be based on several methods rather than on one specific algorithm. This is particularly true in the field of manipulation detection, in which an elaborate forensic analysis with regard to several manipulation techniques obviously increases the quality of the decision whether an image has been tampered with or not. Reliable source identification seems impossible when indications of potential post-processing steps are not taken into account. In a stylized forensic process flow, as depicted in Figure 10, both the manipulation detector and the source identification may consist of one or more forensic algorithms. Information obtained from the manipulation detector may be used to fine-tune the process of source identification, and vice versa. The dashed arrow indicates that a more thorough investigation of image consistency may be feasible once information about the digitization device is available. One step further ahead, one can conceive powerful machine learning classifiers that draw on a large set of orthogonal criteria. Although an attacker might adapt to several forensic algorithms, it is unlikely that he is able to preserve the manifold of all relevant image statistics.

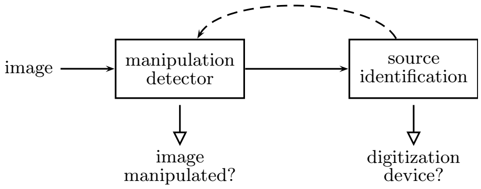

Figure 10: Process flow for a combined image forensic scheme.

Finally, we would like to point out that the two selected forensic techniques are not known to be weak or unreliable tools. On the contrary, these particular methods were chosen in order to build example attacks against powerful and challenging forensic algorithms. And we believe that many other published techniques would be vulnerable to targeted attacks of comparable sophistication. Therefore, the results presented in this paper suggest that the question whether digital image forensics can be trusted has to be answered with no; at least to the extent that currently known techniques are concerned. Whether it will ultimately be possible to construct powerful forensic tools which resist deliberate attacks of informed counterfeiters remains an interesting and — given the number of real-world applications — relevant question for further research.

5. ACKNOWLEDGMENTS

Parts of this work were supported by the German Research Foundation (DFG). Travel to ACM MM 2007 was supported by the European Commission under contract IST-2002-507932 ECRYPT.

6. REFERENCES

[1] W. Chen, Y. Q. Shi, and W. Su. Image splicing detection using 2-D phase congruency and statistical moments of characteristic function. In E. J. Delp and P. W. Wong, editors, *Proc. of SPIE: Security, Steganography, and Watermarking of Multimedia Contents IX*, volume 6072, pages 65050R–1–65050R–8, 2007.

[2] K. S. Choi, E. Y. Lam, and K. Y. Wong. Automatic source camera identification using the intrinsic lens radial distortion. In *Optics Express*, volume 14, pages 11551–11565, Nov. 2006.

[3] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[4] Z. J. Geradts, J. Bijhold, M. Kieft, K. Kurosawa, K. Kuroki, and N. Saitoh. Methods for identification of images acquired with digital cameras. In S. K. Bramblea, E. M. Carapezza, and L. I. Rudin, editors, *Proc. of SPIE: Enabling Technologies for Law Enforcement and Security*, volume 4232, pages 505–512, 2001.

[5] T. Gloe, E. Franz, and A. Winkler. Forensics for flatbed scanners. In E. J. Delp and P. W. Wong, editors, *Proc. of SPIE: Security, Steganography, and Watermarking of Multimedia Contents IX*, volume 6072, pages 65051I–1–65051I–12, 2007.

[6] M. Johnson and H. Farid. Exposing digital forgeries through chromatic aberration. In *Proc. of ACM MM&Sec*, pages 48–55, 2006.

[7] M. Kharrazi, H. T. Sencar, and N. Memon. Blind source camera identification. In *International Conference on Image Processing*, volume 1, pages 709–712, 2004.

[8] M. Kirchner and R. Böhme. Tamper hiding: Defeating image forensics. In *Information Hiding, Ninth International Workshop*, St. Malo, France, June 11.-13. 2007, to appear in LNCS.

[9] Z. Lin, R. Wang, X. Tang, and H.-Y. Shum. Detecting doctored images using camera response normality and consistency. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, volume 1, pages 1087–1092, 2005.

[10] J. Lukáš and J. Fridrich. Estimation of primary quantization matrix in double compressed JPEG images. In *Proc. of the Digital Forensic Research Workshop*, 2003.

[11] J. Lukáš, J. Fridrich, and M. Goljan. Detecting digital image forgeries using sensor pattern noise. In A. Said and J. G. Apostolopoulos, editors, *Proc. of SPIE: Image and Video Communications and Processing 2005*, volume 5685, pages 249–260, 2005.

[12] J. Lukáš, J. Fridrich, and M. Goljan. Digital camera identification from sensor noise. *IEEE Transactions on Information Forensics and Security*, 1(2):205–214, 2006.

[13] M. K. Mıhçak, I. Kozintsev, K. Ramchandran, and P. Moulin. Low-complexity image denoising based on statistical modeling of wavelet coefficients. *IEEE Signal Processing Letters*, 6(12):300–303, Dec. 1999.

[14] T.-T. Ng, S.-F. Chang, C.-Y. Lin, and Q. Sun. Passive-blind image forensics. In W. Zeng, H.Yu, and C.-Y. Lin, editors, *Multimedia Security Technologies for Digital Rights*. Academic Press, 2006.

[15] F. Petitcolas, R. Anderson, and M. Kuhn. Attacks on copyright marking systems. In D. Aucsmith, editor, *Information Hiding, Second International Workshop*, volume LNCS 1525, pages 219–239, Berlin, Heidelberg, 1998. Springer Verlag.

[16] A. Popescu and H. Farid. Exposing digital forgeries by detecting traces of re-sampling. *IEEE Trans. on Signal Processing*, 53(2):758–767, 2005.

[17] A. Popescu and H. Farid. Exposing digital forgeries in color filter array interpolated images. *IEEE Trans. on Signal Processing*, 53(10):3948–3959, 2005.

[18] A. Srivastava, A. Lee, E. Simoncelli, and S.-C. Zhu. On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision*, 18:17–33, 2003.

[19] M. Vrhel, E. Saber, and H. J. Trussell. Color image generation and display technologies. *IEEE Signal Processing Magazine*, pages 23–33, Jan. 2005.

[20] G. Wolberg. *Digital Image Warping*. IEEE Computer Society Press, Los Alamitos, CA, USA, 3. edition, 1994.