Tamper Hiding: Defeating Image Forensics

Matthias Kirchner and Rainer Böhme

Technische Universität Dresden
Institute for System Architecture
01062 Dresden, Germany
matthias.kirchner@acm.org, rainer.boehme@tu-dresden.de

Abstract. This paper introduces novel hiding techniques to counter the detection of image manipulations through forensic analyses. The presented techniques allow to resize and rotate (parts of) bitmap images without leaving a periodic pattern in the local linear predictor coefficients, which has been exploited by prior art to detect traces of manipulation. A quantitative evaluation on a batch of test images proves the proposed method's efficacy, while controlling for key parameters and for the retained image quality compared to conventional linear interpolation.

1 Introduction

Within just one decade, digital signal processing has become the dominant technology for creating, processing and storing the world's pictorial memory. While this new technology clearly has many advantages, critics have expressed concern that it has never been so easy to manipulate images, often in such a perfection that the forgery is visually indistinguishable from authentic photographs. Hence, digitalisation reduces the trustworthiness of pictures in particularly those situations where society is used to base important decisions on them: in the courtroom (photographs as pieces of evidence), in science (published photographs as empirical proofs), and at the ballot box (press photographs).

As a result, research on digital image forensics and tamper detection has gained ground. These techniques can be broadly divided into two branches. One direction tracks particularities of the image acquisition process and reports conspicuous deviations as indications for possible manipulation. Typical representatives of this category include [1,2,3,4,5]. The other approach tries to identify traces from specific image processing functions [6,7,8,9]. Although forensic toolboxes are already quite good at unveiling naive manipulations, they still solve the problem only at its surface. The key question remains open: How reliable are these forensic techniques against a farsighted counterfeiter who is aware of their existence?

To the best of our knowledge, this paper is the first to focus on hiding techniques that help the counterfeiter to defeat forensic tools. We believe that research on "attacks" against forensic techniques is important to evaluate and ultimately improve detectors, as is steganography for steganalysis and vice versa.

Continuing the analogy with steganalysis, one can distinguish *targeted* and *universal* attacks. A targeted attack is a method that avoids traces detectable

T. Furon et al. (Eds.): IH 2007, LNCS 4567, pp. 326–341, 2007.

with one particular forensic technique, which the developer of the attack usually knows. Conversely, universal attacks try to maintain or correct (i.e. make plausible) as many statistical properties of the image to conceal manipulations even when presented to unknown forensic tools. In this sense, a low quality JPEG compression of doctored images can be interpreted as universal attack. While compression often is both plausible and effective—the dominant artefacts from quantisation in the frequency domain are likely to override subtle statistical traces of manipulation—it goes along with a loss in image quality. This highlights the fact that the design space for some attacks against forensic techniques is subject to a trade-off between security (i.e. undetectability) and quality (transparency). This is another parallel to steganography and watermarking.

Tamper hiding techniques can also be classified by their position in the process chain. We call a method *integrated* if it replaces or interacts with the image manipulation operation (e.g. an undetectable copy-move tool as plug-in to image processing software) as opposed to *post-processing*, which refers to algorithms that try to cover all traces after a manipulation with conventional methods.

In this paper we present targeted attacks against a specific technique to detect traces of resampling in uncompressed images proposed by Popescu and Farid [7]. Section 2 recalls the details of this detection method before our countermeasures are discussed in Section 3, together with experimental results. To generalise from single examples and provide a more valid assessment of the proposed methods' performance, a quantitative evaluation on a larger set of test images has been conducted. Its setup and results are given in Section 4. Finally, Section 5 addresses implications for future research on both forensics and counter-forensics.

2 Detecting Traces of Resampling

Most attempts of image forgery rely on scaling and rotation operations, which involve a resampling process. As a result, scholars in image forensics have developed methods to detect traces of resampling in bitmap images. This section reviews the state-of-the-art method proposed by Popescu and Farid [7].

Interpolation algorithms are key to smooth and visually appealing image transformation, however a virtually unavoidable side effect of interpolation is that it introduces linear dependencies between groups of adjacent pixels [10]. The idea of Popescu and Farid's detection method is in identifying these artefacts. They presume that the intensity of each pixel $y_{i,j}$ can be approximated as the weighted sum of pixels in its close neighbourhood (window of size $N \times N$, with $N = 2K + 1$ and K integer) and an independent residual $\boldsymbol{\epsilon}$.

$$y_{i,j} = f(\boldsymbol{\alpha}, \boldsymbol{y}) + \epsilon_{i,j} = \sum_{(k,l)\in\{-K,\ldots,K\}^2} \alpha_{k,l} \cdot y_{i+k,j+l} \; + \; \epsilon_{i,j} \tag{1}$$

They further demonstrate that after interpolation, the degree of dependence from its neighbours differs between pixels. These differences turn out to appear systematically and in a periodic pattern.

The pattern is referred to as p-map and can be obtained from a given image as follows: Using a simplified model, pixels $\boldsymbol{y}$, $y_{i,j} \in [0, 255]$, are assigned to one of two classes $\mathcal{M}_1$ and $\mathcal{M}_2$. Set $\mathcal{M}_1$ contains those pixels with high linear dependence whereas set $\mathcal{M}_2$ comprises all pixels without it. The *expectation maximisation* (EM) algorithm [11], an iterative two-stage procedure, allows to estimate simultaneously both, the set a specific pixel most likely belongs to, and the unknown weights $\boldsymbol{\alpha}$. First, the E-step uses the Bayes theorem to calculate the probability for each pixel belonging to set $\mathcal{M}_1$.

$$p_{i,j} = \text{Prob}(y_{i,j} \in \mathcal{M}_1 | y_{i,j}) = \frac{\text{Prob}(y_{i,j} | y_{i,j} \in \mathcal{M}_1) \cdot \text{Prob}(y_{i,j} \in \mathcal{M}_1)}{\sum_{k=1}^{2} \text{Prob}(y_{i,j} | y_{i,j} \in \mathcal{M}_k) \cdot \text{Prob}(y_{i,j} \in \mathcal{M}_k)} \tag{2}$$

Evaluating this expression requires

1. a conditional distribution assumption for $\boldsymbol{y}$: $y \sim \mathcal{N}(f(\boldsymbol{\alpha}, \boldsymbol{y}), \sigma_{\mathcal{M}_1})$ for $y_{i,j} \in \mathcal{M}_1$ and $y \sim \mathcal{U}(0, 255)$ for $y_{i,j} \in \mathcal{M}_2$,
2. knowledge of weights $\boldsymbol{\alpha}$ (initialised with $1/(N^2 - 1)$ in the first round),
3. knowledge of $\sigma_{\mathcal{M}_1}$ (initialised with the signal's empirical standard deviation),
4. another assumption saying $\text{Prob}(y_{i,j} \in \mathcal{M}_1) = \text{Prob}(y_{i,j} \in \mathcal{M}_2)$.

In the M-step, vector $\boldsymbol{\alpha}$ is updated using a weighted least squares estimator:

$$\boldsymbol{\alpha} = (\boldsymbol{Y}'\boldsymbol{W}\boldsymbol{Y})^{-1}\boldsymbol{Y}'\boldsymbol{W} \cdot \boldsymbol{y} \tag{3}$$

Matrix $\boldsymbol{Y}$ has dimension $|\boldsymbol{y}| \times (N^2 - 1)$ and contains the non-center elements of all windows as stacked row vectors. Matrix $\boldsymbol{W}$ holds the corresponding conditional probabilities $p_{i,j}$ of (2) as weights on its diagonal, hence $\boldsymbol{p} = \text{diag}(\boldsymbol{W})$. Given new estimates for $\boldsymbol{p}$ and $\boldsymbol{\alpha}$, $\sigma_{\mathcal{M}_1}$ can be computed as weighted standard deviation from the residuals $\boldsymbol{\epsilon}$. E-step and M-step are iterated until convergence.

Previous resampling operations leave periodical pattern in the so-obtained p-maps. This pattern becomes most evident after a transformation into the frequency domain, using a Discrete Fourier Transformation (DFT), where it causes distinct peaks that are typical for the specific resampling parameters. To enhance the visibility of the characteristic peaks, Popescu and Farid propose to apply a contrast function C [7]. The contrast function is composed of a radial weighting window, which attenuates very low frequencies, and a gamma correction step. The absolute values of the resulting complex plane can be visualised and presented to a human forensic investigator.

Figure 1 illustrates the detection process by comparing an original greyscale image to a processed version that has been scaled up[1] with linear interpolation to 105 % of the original (left column). The resulting p-maps are displayed in the centre. As expected, the rather chaotic p-map of the original image shows a very

[1] We show upscaling because it is particularly likely to leave detectable traces in the redundancy of newly inserted pixels. So it forms a critical test for our methods.

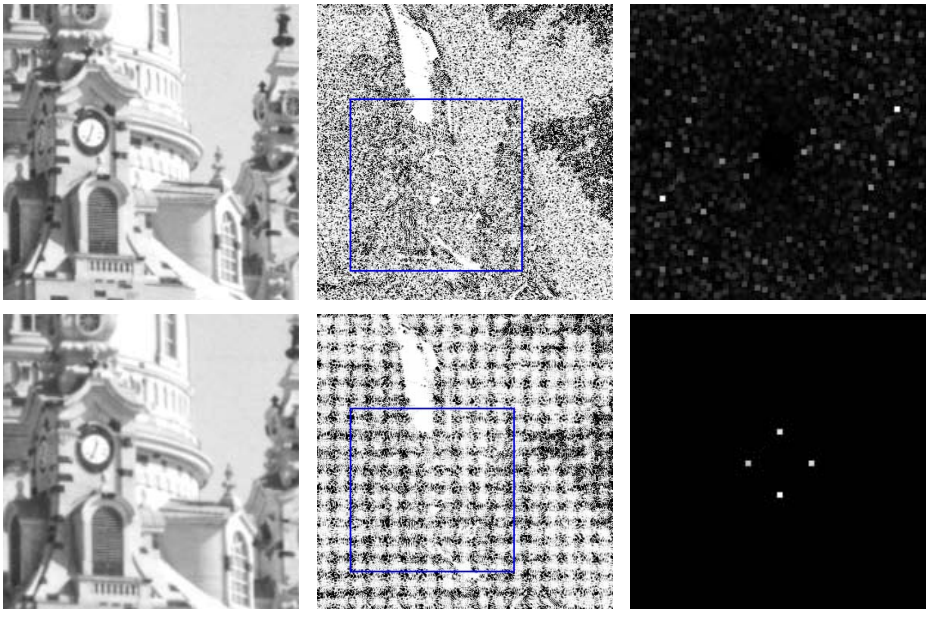

Fig. 1. Results of resampling detection for original image (top row) and 5 % upsampling (bottom row). Complete p-maps are displayed in the centre column; frames mark the parts depicted on the left. Periodic resampling artefacts lead to characteristic peaks in the corresponding spectrum (rightmost pictures).

clear periodic structure after transformation, which also explains the different appearance of the spectrum (right column). To enhance the quality in print, each spectrum graph in this paper is normalized to span the full intensity range. We further apply a maximum filter to improve the visibility of the peaks.

In general, this detection method is known as an effective and powerful tool. Robustness against several image manipulation operations (except lossy compression) has already been proven in the original publication and could be confirmed by us, also with respect to non-linear interpolation methods, such as B-splines.

3 Countermeasures Against Resampling Detection

In the hand of forensic investigators, this powerful detection method might raise the temptation to use its results as proof of evidence in legal, social and scientific contexts. However, one must bear in mind that forensic methods merely provide indications and are by orders of magnitude less dependable than other techniques, such as decent cryptographic authentication schemes. In contrast to cryptography, multimedia forensics remains an inexact science without rigourous security proofs. To draw attention to this problem, we will present three methods to perform image transformations that are almost undetectable by the above

described method. In this sense, these techniques can be considered as attacks against the detection algorithm.

3.1 Attacks Based on Non-linear Filters

The detection method is based on the assumption of systematic linear dependencies between pixels in close neighbourhood (see Eq. (1)). Hence, all kinds of non-linear filters, applied as post-processing step, are candidates for possible attacks. The *median filter*, a frequently used primitive in image processing [12], replaces each pixel with the median of all pixels in a surrounding window of defined shape and size. This acts as a low-pass filter, however with floating cutoff frequency.

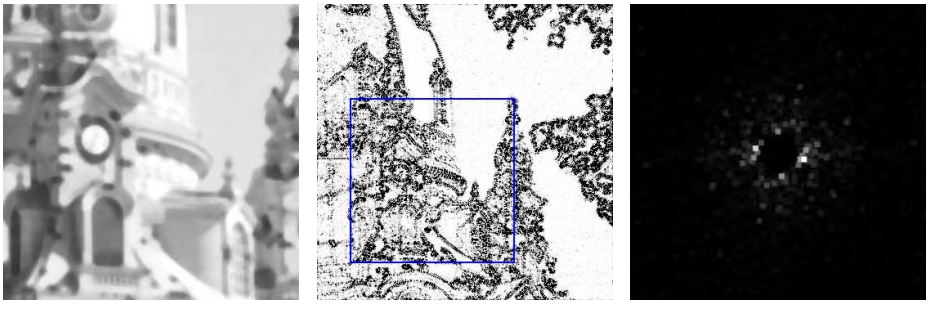

Fig. 2. Results after upsampling by 5 % and post-processing with a 5×5 median filter: characteristic peaks in the spectrum vanish, however the image appears excessively blurred

Figure 2 shows the results of the detection algorithm applied on a transformed image that has been post-processed with a 5×5 square median filter. This attack is successful as the characteristic peaks in the spectrum have disappeared. Note that the amplitudes corresponding to the brightest spots in the rightmost graph are by magnitudes smaller than the peaks in Fig. 1. However, a simple median filter negatively affects the quality of the post-processed image, which is reflected in noticeable blurring. Therefore, despite effective, naive non-linear filters are suboptimal for mounting relevant attacks in practice.

3.2 Attacks Based on Geometric Distortion

Inspired by the effectiveness of geometric attacks against watermarking schemes [13], we have explored geometric distortion as building blocks for attacks against tamper detection. We expect it to be effective in our application as well because the detection method exploits the periodic structure in mapping discrete lattice position from source to destination image, where the relative position of source and target pixels is repeated over the entire plane. This systematic similarity allows to separate it statistically from residual image content. To break

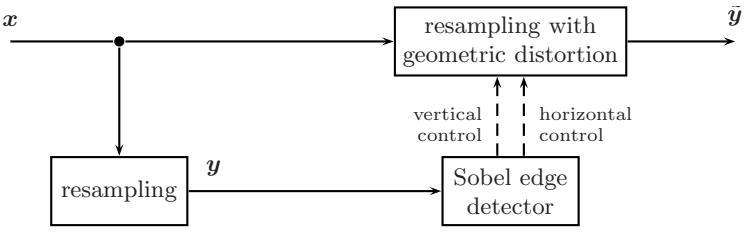

Fig. 3. Block diagram of geometric distortion with edge modulation

the similarity, each individual pixel's target position is computed from the transformation relation with a random disturbance vector $\boldsymbol{e}$ superimposed.

$$\begin{bmatrix} i \\ j \end{bmatrix} = \boldsymbol{A} \cdot \begin{bmatrix} i_{\boldsymbol{x}} \\ j_{\boldsymbol{x}} \end{bmatrix} + \begin{bmatrix} e_{1,i,j} \\ e_{2,i,j} \end{bmatrix} \quad \text{where} \quad e \sim \mathcal{N}(0,\sigma) \quad \text{i. i. d.} \tag{4}$$

$\boldsymbol{A}$ is the transformation matrix and indices $i_{\boldsymbol{x}}, j_{\boldsymbol{x}}$ refer to source positions as opposed to i, j which index the resampled image. Parameter σ controls the degree of distortion. However, naive geometric distortion may cause visible artefacts, such as jitter, which is perceived most visually disturbing at straight lines and edges. To evade such quality loss, we modulate the strength of distortion adaptively from the local image content. The modulation is controlled by two edge detectors, one for horizontal and one for vertical disturbance, as follows:

$$\begin{bmatrix} i \\ j \end{bmatrix} = \boldsymbol{A} \cdot \begin{bmatrix} i_{\boldsymbol{x}} \\ j_{\boldsymbol{x}} \end{bmatrix} + \begin{bmatrix} e_{1,i,j} \cdot (1 - 1/255 \cdot \text{sobelH}(\boldsymbol{y}, i_{\boldsymbol{y}}, j_{\boldsymbol{y}})) \\ e_{2,i,j} \cdot (1 - 1/255 \cdot \text{sobelV}(\boldsymbol{y}, i_{\boldsymbol{y}}, j_{\boldsymbol{y}})) \end{bmatrix} \quad . \tag{5}$$

Functions sobelH and sobelV return the value of a linear Sobel filter for horizontal and vertical edge detection, respectively [12]. This construction applies fewer distortion to areas with sharp edges, where the visible impact would be most harmful otherwise. The Sobel filter coefficients are defined as

$$\boldsymbol{H} = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{V} = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad .$$

Our implementation ensures that the range is truncated to the interval $[0, 255]$. Note that the filter is applied to a transformed image without any distortion $\boldsymbol{y}$. As a consequence, this attack requires the image to be transformed twice, as depicted in the block diagram of Fig. 3.

The results demonstrate that geometric distortion is capable to eliminate the characteristic traces from the p-map spectrum (Fig. 4). In line with our expectations, the edge modulation mitigates the loss in image quality considerably.

3.3 A Dual Path Approach to Undetectable Resampling

While geometric distortion with edge modulation generates already good results, we found from a comprehensive evaluation of many different transformation pa-

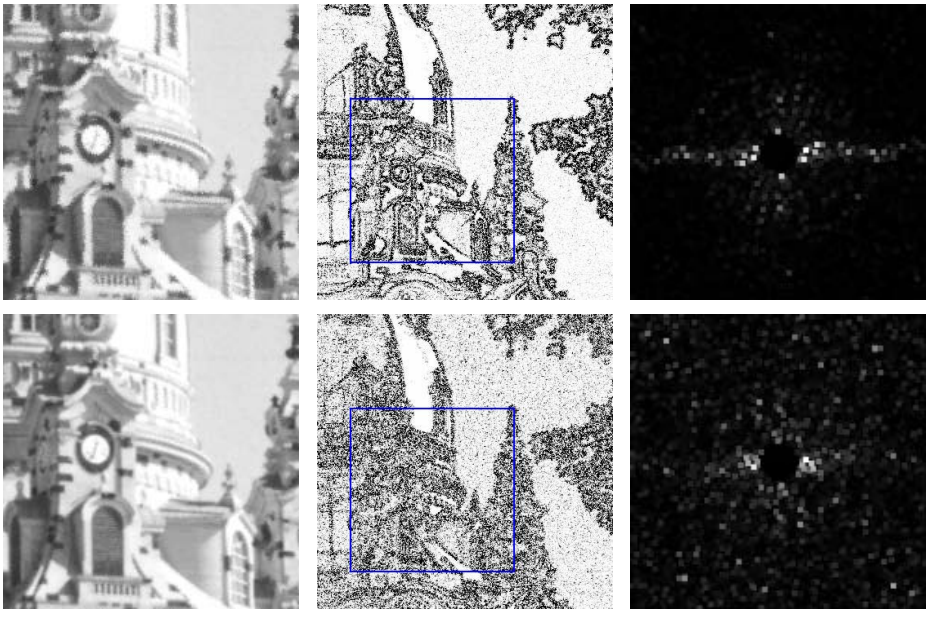

Fig. 4. Results after upsampling by 5 % with geometric distortion of strength $\sigma = 0.4$. Comparison between naive distortion (top) and edge modulation using horizontal and vertical Sobel filters (bottom).

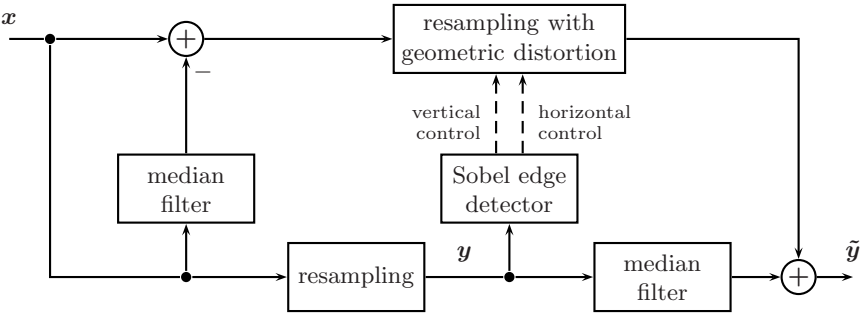

Fig. 5. Block diagram of dual path approach: combination of median filter for low frequency image component and geometric distortion with edge modulation for the high frequency component

rameters that the undetectability can be improved further by applying different operations to the high and low frequency components of the image signal. Such approaches have already been applied successfully in noise reduction [14] and watermarking attacks [15]. Figure 5 illustrates the proposed process. The two frequency components are separated with a median filter. First, the low frequency component of the output image is obtained by applying a median filter

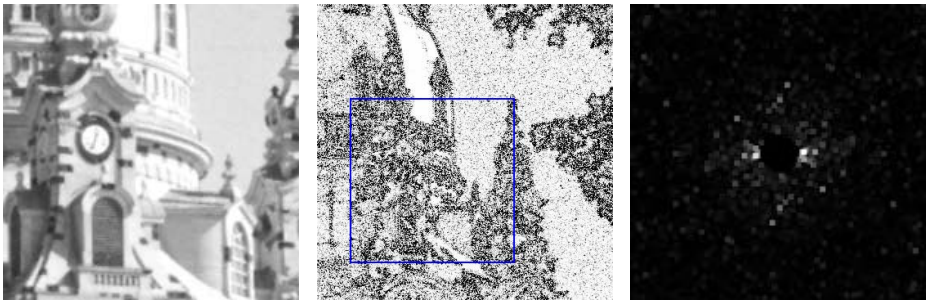

Fig. 6. Dual path method: 5 % upsampling, 7×7 median filter for low frequency component combined with geometric distortion ($\sigma = 0.3$) and edge modulation

directly to the resampled source image (see Sect. 3.1). Second, a high frequency component is extracted from the source image $\boldsymbol{x}$ by subtracting the result of a median filter (other low-pass filters are conceivable as well). This component is resampled with geometric distortion and edge modulation (see Sect. 3.2), where the edge information is obtained from the resampled image $\boldsymbol{y}$ prior to the median filter. The final image $\tilde{\boldsymbol{y}}$ is computed by summing up both components. This attack has two parameters, the size of the median filter and the standard deviation of the geometric distortion σ.

Figure 6 finally reports the results of the dual path approach. It becomes evident that the obtained p-map is most similar to the p-map of the original (see Fig. 1 above). Further, no suspicious peaks appear in its spectrum. The image quality is preserved and shows no visible artefacts.

4 Quantitative Evaluation

For a quantitative evaluation of our attacks against resampling detection, we built a database of 168 never-compressed 8 bit greyscale images, each of dimension 426×426 pixels. All images were derived from a smaller set of 14 photographs taken with a Nikon Coolpix 4300 digital camera at full resolution (2272×1704). Therefore we first cut every photograph into twelve 852×852 parts with maximum 50 % overlap. Then each part was downsampled by factor two to avoid possible interference from periodic patterns that might stem from a *colour filter array* (CFA) interpolation inside the camera [2].

As described in Sect. 2, the resampling detector relies on finding periodic dependencies between pixels in a close neighbourhood. To identify forgeries automatically, Popescu and Farid propose to measure the similarity between the p-map of a given image and a set of synthetically generated periodic patterns [7]. The synthetic map $\boldsymbol{s}^{(\boldsymbol{A})}$ for transformation $\boldsymbol{A}$ is generated by computing the distance between each point in the resampled lattice and the closest point in the original lattice,

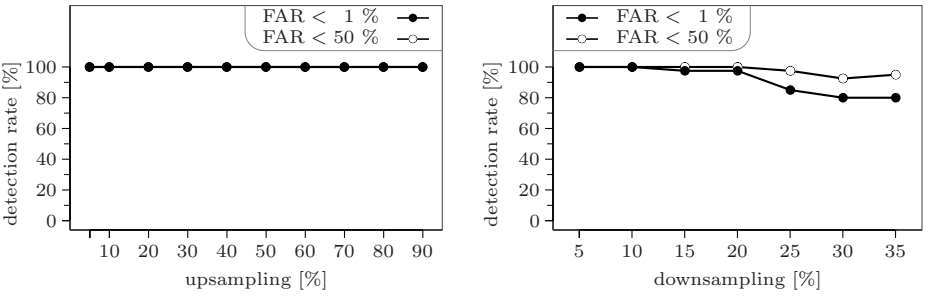

Fig. 7. Results of resampling detection after upsampling (left) and downsampling (right) by varying amounts. Each data point corresponds to the resampling of 40 images.

$$s_{i,j}^{(\boldsymbol{A})} = \left\| \boldsymbol{A} \cdot \begin{bmatrix} i \\ j \end{bmatrix} - \left\lfloor \boldsymbol{A} \cdot \begin{bmatrix} i \\ j \end{bmatrix} + \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix} \right\rfloor \right\| \quad . \tag{6}$$

In the absence of prior information about the actual transformations parameters $\boldsymbol{A}$, an automatic detector conducts an exhaustive search in a set $\mathcal{A}$ of candidate transformation matrices $\boldsymbol{A}_q$. In all our experiments, $\mathcal{A}$ contains 256 synthetic maps for upsampling in the range of 1 % to 100 % as well as 128 synthetic maps for downsampling in the range of 1 % to 50 % using equidistant steps of 0.4 percentage points. The maximum pairwise similarity between an empirical p-map and all elements of $\mathcal{A}$ is taken as a decision criterion d.

$$d = \max_{\boldsymbol{A} \in \mathcal{A}} \sum_{i,j} \left| \mathrm{C}(\mathrm{DFT}(\boldsymbol{p})) \right| \cdot \left| \mathrm{DFT}\left(\boldsymbol{s}^{(\boldsymbol{A})} \right) \right| \tag{7}$$

Function C is the contrast function (see above) and DFT applies a 2D discrete Fourier transformation. If d exceeds a specific threshold d_T then the corresponding image is flagged as resampled. We have determined d_T empirically for a defined *false acceptance rate* (FAR) by applying the detector to all 168 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.

Figure 7 reports the **baseline detection results** for upsampling and downsampling using plain linear interpolation. Each data point is computed as average from 40 resampled images.[2] We find perfect detection for upsampling and very high detection accuracy for downsampling. This confirms the general effectiveness of the detection method in the range of tested transformation parameters. Thus, Figure 7 may serve as reference for the evaluation of our attacks with respect to their capability to hide such image transformations.

[2] The detector parameters were set to $N = 2$ and $\|\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_{n-1}\| < 0.001$ as convergence criterion for the EM algorithm. The modest amount of images is due to the computational complexity of about 50 seconds computation time for one single p-map using a C implementation on a 1.5 Ghz G4 processor.

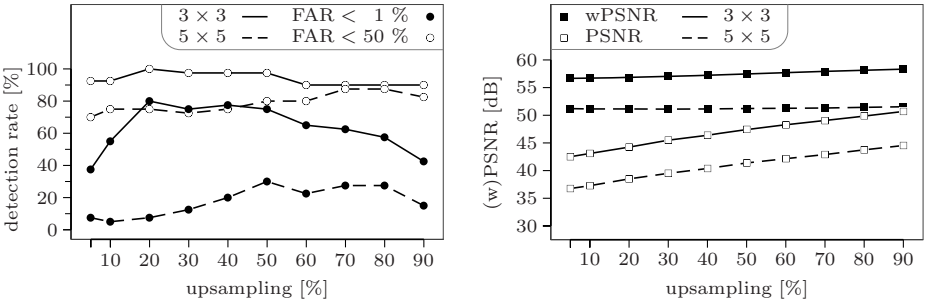

Fig. 8. Evaluation of median filter at different window sizes. Detection rates (left) and average image quality (right). Larger window sizes reduce both detection rates and image quality.

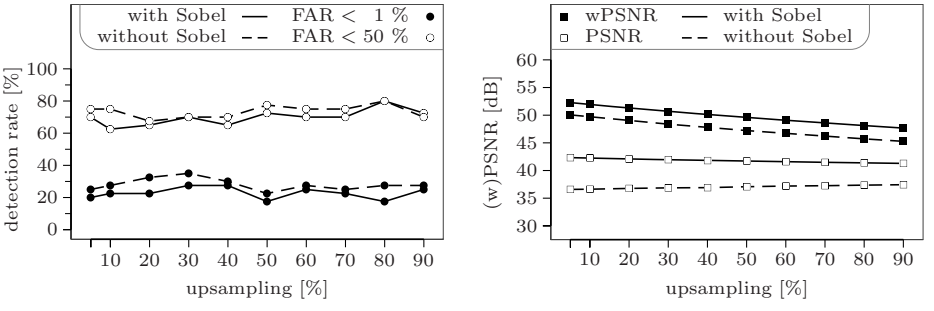

Fig. 9. Evaluation of geometric distortion ($\sigma = 0.4$) with and without edge modulation. Detection rates (left) and image quality (right). Edge modulation yields substantially better quality and slightly superior detection results.

Any attempt to conceal resampling operations should not only be judged by the achieved level of undetectability but also by the amount of image degradation. For our quantitative evaluation we resort to common image quality metrics Q to assess the visual impact of our proposed attacks.

$$Q = 20 \ \log \frac{255}{\|(\boldsymbol{y} - \tilde{\boldsymbol{y}}) \cdot \boldsymbol{v}\|} \tag{8}$$

We report the metrics PSNR, where $\boldsymbol{v} = \mathbf{1}$, as well as a variant adjusted for human visual perception wPSNR (*'w'* for *weighted*). It has been argued that the latter metric is a more valid indicator for the evaluation of watermarking attacks [16]. Weights $\boldsymbol{v}$ are computed from a *noise visibility function* (NVF), which emphasises image regions with high local variance and attenuates flat regions and soft gradients. Among the two NVFs proposed in [17] we have chosen the one based on a stationary Generalised Gaussian image model. Both metrics are measured in dB. Higher values indicate superior image quality.

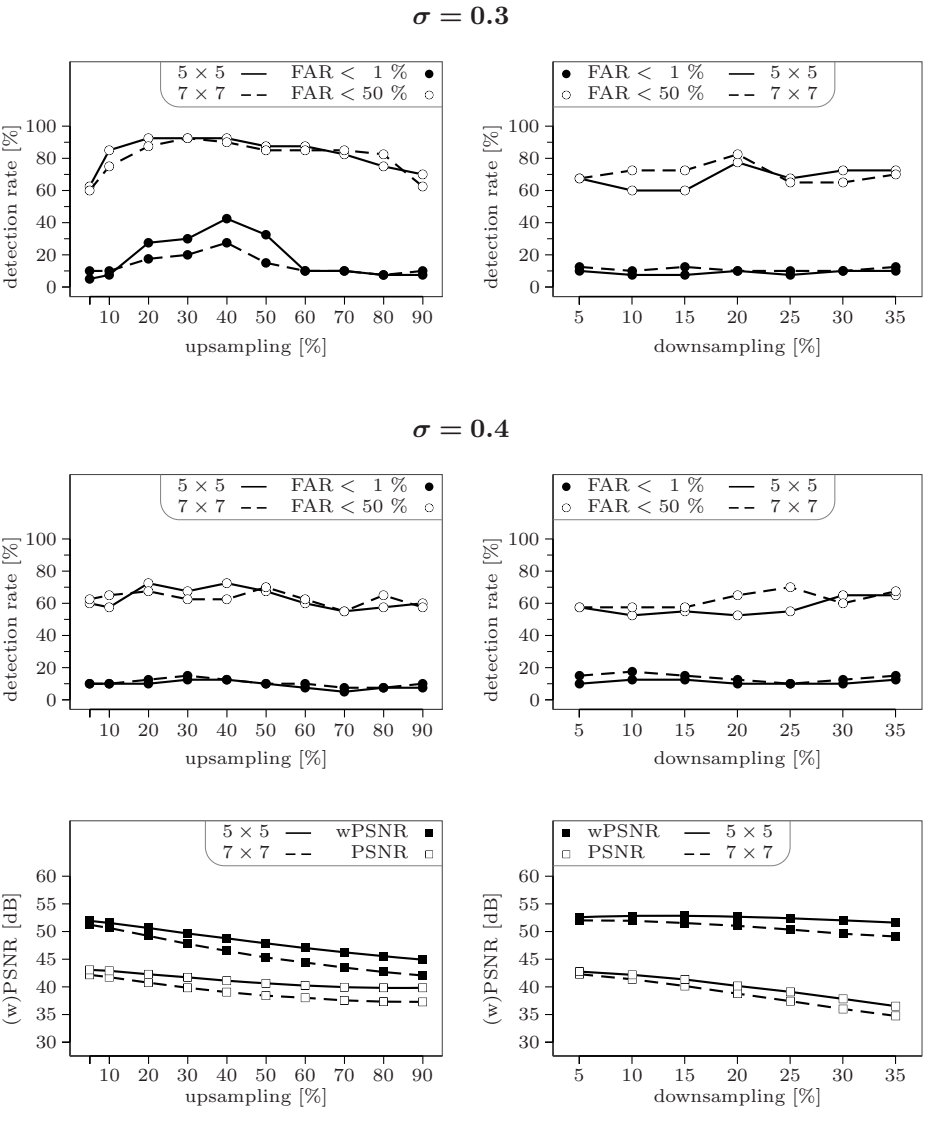

Fig. 10. Evaluation of dual path approach for upsampling (left column) and downsampling (right column). Detection rates for $\sigma = 0.3$ (top row) and $\sigma = 0.4$ (centre row); average image quality for $\sigma = 0.4$ (bottom row). Breakdown by window size of the median filter (5×5 vs 7×7) and false acceptance rates (FAR: 1 % vs 50 %). Stronger distortion in the high frequency component decreases detectability. Smaller windows sizes in the low frequency component retain better image quality.

Figure 8 reports detection rates (left) and average image quality (right) for upsampled images, post-processed with **median filters** of sizes 3×3 and 5×5, respectively. As larger window sizes introduce a higher degree of non-linearity,

the 5×5 median filter yields noticeable less detectable results than the smaller 3×3 filter. However, this comes at a cost of substantial losses in visual image quality, which can be expressed both in terms of PSNR and wPSNR.

Note that the success of this attack depends on the upsampling ratio in a non-linear manner. The results for downsampling are omitted for the sake of brevity.

Since, at practical window sizes, median-filtered images suffer from extensive blurring, we have further investigated the effect of **geometric distortion** in the resampling process. Figure 9 shows the results for upsampling by varying amounts with distortion of strength $\sigma = 0.4$. As can be seen from the graphs, using edge modulation is a reasonable extension to the general approach. While detection rates remain stable on a relatively low level for all tested transformation parameters both with and without edge modulation, the latter yields a considerable improvement in image quality between 2–6 dB on average.

Finally, Figure 10 presents the results for the **dual path approach**. Since we consider this method as benchmark for future research, graphs for both upsampling (left column) and downsampling (right column) are displayed. The top four charts show detection rates for distortion strengths $\sigma = 0.3$ and $\sigma = 0.4$, respectively. Average image quality for $\sigma = 0.4$ is reported in the bottom row. The frequency components have been separated with 5×5 and 7×7 median filters. While a higher degree of geometric distortion generally reduces detection rates, we found that the choice of σ is more important for upsampling than for downsampling. Note that both 5×5 and 7×7 median filter lead to similar detection rates, however the former might be preferred with regard to image quality metrics. A direct comparison of the dual path approach with geometric distortion as described in Sect. 3.2 (Fig. 9) reveals a clear advantage of the dual path approach. For $\sigma = 0.4$, the latter achieves considerably better undetectability whereas image quality metrics indicate only marginal losses.

The very low detection rates of the dual path approach for $\sigma = 0.4$ demonstrate how successfully resampling operations can be concealed with the proposed method. At a practically relevant false acceptance rate $< 1\,\%$, only about $10\,\%$ of all image transformations were correctly identified as resampled (5×5 median filter, $\sigma = 0.4$). To allow for a better comparability with future research, detailed numeric results including summary statistics for the decision criterion d are given in Table 2 in the appendix. We further found that the few successful detections were concentrated within just a couple of original images, which suggests that image-specific factors may determine the efficacy of our attack.

Note that we have also tested the robustness of our results for detectors with smaller ($N = 1$) and larger ($N = 3$) neighbourhoods. As the corresponding dual path detection rates do not differ substantially from the reported figures, we conclude that our results are fairly robust and refrain from reporting them separately.

5 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. In particular, we

have presented and evaluated three approaches to defeat a specific method of resampling detection, which has been developed to unveil scaling and rotation operations of digital images or parts thereof. These attacks have turned out to be the most effective ones in a broader research effort, which also led to a number of dead ends. Table 1 in the appendix briefly documents our less successful attempts as guidelines for future research in the area. Among the successful methods, the dual path approach, which applies geometric distortion with edge modulation to the high frequency component of an image signal and a median filter to the (low frequency) residual, achieved the best performance and should be regarded as benchmark for other specific tamper hiding techniques. At the same time, we would like to point out that the resampling detector of Popescu and Farid [7], against which our work in this paper is targeted, is certainly not a weak or unreliable tool when applied to plain interpolation. On the contrary, we have selected this particular detector with the aim to build an example attack against a powerful and challenging method. And we believe that many other published techniques would be vulnerable to targeted attacks of comparable sophistication.

Apart from the detailed results presented in the previous section, there are at least two more general conclusions worth mentioning. First, attacks which are integrated in the manipulation operation appear to be more effective than others that work at a post-processing step. This is plausible, since information about the concrete transformation parameters is not available at the post-processing stage and therefore much stronger interference with the image structure is necessary to cover up statistical artefacts of all possible transformations in general. Second, a closer look at all quantitative results suggests that it is easier to conceal downscaling than upscaling. This is plausible as well, since downscaling causes information loss, whereas it is more difficult to impute new pixels with idiosyncratic information. This implies that larger window sizes (for the median filter approach) and stronger geometric distortion are necessary for upscaling to achieve similar levels of (un)detectability as for downscaling.

As to the limitations, we consider this work as a first and modest attempt in an interesting sub-field. It is obvious that our results hold only for the specific detection method and we cannot rule out that image manipulations conducted with our proposed methods are detectable with a) other existing forensic techniques or b) new targeted detection methods that are build with the intention to discover our attacks. While this might trigger an new cat-and-mouse race between forensic and counter-forensic techniques, we believe that such creative competition is fruitful and contributes to a more holistic picture on the possibilities and limitations of image forensics, an area where much prior research has been done against the backdrop of a fairly naive ‘adversary model’—a term borrowed from cryptography, where dealing with strong adversaries has a longer tradition [18]. On a more abstract level, one may ask the question whether it is possible at all to construct provable secure techniques under gentle assumptions. We conjecture that an ultimate response is far distant and it is probably

linked to related questions, such as the search for provable secure high capacity steganography (with realistic cover assumptions), and to the development of better stochastic image models. In the meantime, more specific research questions are abundant.

Acknowledgements

The first author gratefully acknowledges receipt of a student travel grant awarded by Fondation Michel Métivier, France.

References

1. Ng, T.-T., Chang, S.-F.: A Model for Image Splicing. In: Proc. of ICIP 2004, vol. 2, pp. 1169–1172 (2004)
2. Popescu, A., Farid, H.: Exposing Digital Forgeries in Color Filter Array Interpolated Images. IEEE Trans. on Signal Processing 53, 3948–3959 (2005)
3. Lukáš, J., Fridrich, J., Goljan, M.: Detecting Digital Image Forgeries Using Sensor Pattern Noise. In: Delp, E.J., Wong, P.W. (eds.) Proc. of SPIE: Security and Watermarking of Multimedia Content VII, vol. 6072, pp. 60720Y-1–60720Y-11 (2006)
4. Johnson, M., Farid, H.: Exposing Digital Forgeries through Chromatic Aberration. In: Proc. of ACM MM-Sec., pp. 48–55 (2006)
5. Swaminathan, A., Wu, M., Liu, K.: Image Tampering Identification Using Blind Deconvolution. In: Proc. of ICIP 2006, pp. 2311–2314 (2006)
6. Fridrich, J., Soukal, D., Lukáš, J.: Detection of Copy-Move Forgery in Digital Images. In: Proc. of the Digital Forensic Research Workshop (2003)
7. Popescu, A.C., Farid, H.: Exposing Digital Forgeries by Detecting Traces of Resampling. IEEE Trans. on Signal Processing 53, 758–767 (2005)
8. Johnson, M.K., Farid, H.: Exposing Digital Forgeries by Detecting Inconsistencies in Lighting. In: Proc. of ACM MM-Sec., pp. 1–10 (2005)
9. Farid, H.: Exposing Digital Forgeries in Scientific Images. In: Proc. of ACM MM-Sec. pp. 29–36 (2006)
10. Thévenaz, P., Blu, T., Unser, M.: Interpolation Revisited. IEEE Trans. on Medical Imaging 19, 739–758 (2000)
11. Dempster, A.P., Laird, N.M., Rubin, D.: Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B 39, 1–38 (1977)
12. Pitas, I.: Digital Image Processing Algorithms and Applications. John Wiley & Sons, Inc., Chichester (2000)
13. Petitcolas, F., Anderson, R., Kuhn, M.: Attacks on Copyright Marking Systems. In: Aucsmith, D. (ed.) IH 1998. LNCS, vol. 1525, pp. 219–239. Springer, Heidelberg (1998)
14. Bernstein, R.: Adaptive Nonlinear Filters for Simultaneous Removal of Different Kinds of Noise in Images. IEEE Trans. on Circuits and Systems 34, 1275–1291 (1987)
15. Langelaar, G.C., Biemond, J., Lagendijk, R.L.: Removing Spatial Spread Spectrum Watermarks by Non-Linear Filtering. In: Proc. of EUSIPCO 1998, pp. 2281–2284 (1998)

16. Voloshynovskiy, S., Pereira, S., Herrigel, A., Baumgaertner, N., Pun, T.: Generalized Watermarking Attack Based on Watermark Estimation and Perceptual Remodulation. In: Wong, P.W., Delp, E.J. (eds.) Proc. of SPIE: Security and Watermarking of Multimedia Content II, vol. 3971, pp. 358–370 (2000)
17. Voloshynovskiy, S., Herrigel, A., Baumgaertner, N., Pun, T.: A stochastic Approach to Content Adaptive Digital Image Watermarking. In: Pfitzmann, A. (ed.) IH 1999. LNCS, vol. 1768, pp. 212–236. Springer, Heidelberg (2000)
18. Kerckhoffs, A.: La cryptographie militaire. Journal des sciences militaires IX, 5–38, 161–191 (1883)

Appendix

Table 1. Summary of alternative attack methods investigated in the literature and in the course of this research

Method	Type[a)]	Success[b)]	Image quality[c)]
Existing literature [7]			
Additive noise	P	−	−
Gamma correction	P	−	−
JPEG compression	P	+	+
JPEG2000 compression	P	∘	+
Our research			
Mean filter	P	−	−
Binomial filter	P	−	−
Multistage median filter	P	−	∘
Incremental resampling 1	P	−	−
Incremental resampling 2	I	+	−
Locally correlated geometric distortion	I	−	+
Dual path with extremum filter (HF)	P	−	∘

[a)] I integrated, P post-processing

[b)] + manipulation undetectable, − manipulation detectable, ∘ parameter dependent

[c)] + good quality (only plausible artefacts), − visible distortion, ∘ parameter dependent

Table 2. Detailed results for dual path approach ($\sigma = 0.4$, window size 5×5)

	d		detection rate [%]		average image quality [a]	
	median	IQR [b]	FAR < 1 %	FAR < 50 %	wPSNR [dB]	PSNR [dB]
Originals (168 images)						
	20.32	37.20	–	–	–	–
Upsampling (40 images each)						
5 %	29.74	91.97	10.0	60.0	51.89 (1.65)	43.03 (4.71)
10 %	32.60	101.54	10.0	57.5	51.51 (1.76)	42.86 (4.77)
20 %	39.88	82.96	10.0	72.5	50.60 (1.93)	42.24 (4.94)
30 %	32.04	58.53	12.5	67.5	49.61 (2.10)	41.68 (5.03)
40 %	32.80	40.14	12.5	72.5	48.73 (2.23)	41.07 (5.14)
50 %	29.35	84.09	10.0	67.5	47.82 (2.39)	40.60 (5.26)
60 %	28.26	65.28	7.5	60.0	47.00 (2.65)	40.22 (5.50)
70 %	27.20	62.12	5.0	55.0	46.21 (2.82)	39.88 (5.51)
80 %	27.10	56.06	7.5	57.5	45.50 (3.00)	39.80 (5.51)
90 %	29.06	46.01	7.5	60.0	44.88 (3.40)	39.79 (5.77)
average detection rate [a]			**9.3** (2.4)	**63.0** (6.4)		
Downsampling (40 images each)						
5 %	23.80	106.98	10.0	57.5	52.54 (1.62)	42.72 (4.69)
10 %	23.89	100.14	12.5	52.5	52.72 (1.66)	42.13 (4.79)
15 %	24.44	84.18	12.5	55.0	52.78 (1.63)	41.29 (4.85)
20 %	25.57	78.95	10.0	52.5	52.67 (1.74)	40.19 (4.94)
25 %	39.67	84.89	10.0	55.0	52.37 (1.87)	39.06 (4.98)
30 %	39.85	96.21	10.0	65.0	51.97 (2.06)	37.79 (5.04)
35 %	48.54	85.13	12.5	65.0	51.57 (2.07)	36.46 (5.03)
average detection rate [a]			**11.1** (1.3)	**57.5** (5.4)		

[a] standard deviation in brackets [b] inter-quartile range (measure of dispersion)