PRNU-based Image Manipulation Localization with Discriminative Random Fields

Sujoy Chakraborty and Matthias Kirchner
Department of Electrical and Computer Engineering
Binghamton University; Binghamton, NY, USA

Abstract

We formulate PRNU-based image manipulation localization as a probabilistic binary labeling task in a flexible discriminative random field (DRF) framework. A novel local discriminator based on the deviation of the measured correlation from the expected local correlation as estimated by a correlation predictor is paired with an explicit pairwise model for dependencies between local decisions. Experimental results from the Dresden Image Database indicate that the DRF outperforms prior art with Markov random field label priors.

Introduction

Over the last decade, scholars and practioners have embraced digital camera sensor noise as one of the most valuable image characteristics in digital image forensics [1]. A wide range of works pioneered by Fridrich et al. [2] indicate that virtually all digital cameras leave a unique camera-specific sensor noise fingerprint in the images they capture. Minute manufacturing imperfections of individual sensor elements lead to a spatially varying multiplicative noise pattern, the photo-response non-uniformity (PRNU), that can be estimated and tested for in forensic applications. This is particularly useful for source attribution, where the goal is to link digital images to their source camera. Another application is the detection and localization of local image manipulations. When content in an image is replaced with content from elsewhere, or when certain areas undergo strong processing, those non-genuine regions will lack the expected local portion of the camera fingerprint. A straightforward manipulation detection algorithm inspects the query image in small analysis windows and compares the local noise estimate to the corresponding part of the camera fingerprint. The content in the analysis window is likely not genuine, if the correlation-based similarity score falls below a suitably chosen threshold.

A general disadvantage of sliding-window image manipulation detection algorithms is that local decisions are made independent of their surroundings. Specifically, local outcomes are likely to depend on each other when a manipulation spans multiple analysis windows. More generally, local decisions may also be influenced by non-local image characteristics (e. g., global brightness or noise level). Most current forensic schemes do not incorporate such contextual information explicitly. The work by Chierchia et al. [3] is a recent exception. The authors frame the problem of manipulation localization as a probabilistic labeling task that maps local analysis neighborhoods to a binary label field. A rigorous Bayesian Markov random field model was introduced as a smoothness prior to account for slowly varying label characteristics.

Along these lines, our goal is to infer local label assignments y_i from a global posterior $p(\mathbf{y}|\mathbf{x})$, which expresses how likely a choice of labels $\mathbf{y}$ is, given the observed data $\mathbf{x}$. The major challenge here is to find meaningful yet tractable models of $p(\mathbf{y}|\mathbf{x})$. We draw on a discriminative random field (DRF) formulation [4] for this purpose, which differs from Chierchia et al.'s work [3] mainly in that i) a DRF model is built around local discriminative classifiers (as opposed to generative models), and ii) a DRF allows label dependencies to be data-dependent. Instead of assuming a homogenous label interaction prior $p(\mathbf{y})$ with Markov properties (for instance the Ising model), a discriminative random field directly models the posterior $p(\mathbf{y}|\mathbf{x})$ as being Markov. The result is a more flexible model that permits interaction in both the observed data and the labels in a principled manner [4]. Before we elaborate on these aspects in more detail below, the next section provides a brief review of PRNU-based image forensics. We then describe the mathematical foundation of the DRF framework, discuss our specific instantiation, followed by our experimental setup, results and conclusion.

Notation For notational convenience, we represent matrices of size $U \times V$ as UV-vectors, thereby keeping spatial structure implicit. Vector indices $i \in \mathcal{S} = \{0, 1, \ldots, UV-1\}$ will also be referred to as *sites*. Specifically, we denote grayscale images as $\mathbf{x} \in \mathcal{X} = \{0, 1, \ldots, 255\}^{|\mathcal{S}|}$, with x_i being the i-th pixel in column-major order. As we will be concerned with quantities extracted from analysis windows of size $W \times W$, possibly overlapping by $0 \leq O < W$ elements in horizontal and vertical direction, denote the m-th *analysis window* in an $U \times V$ matrix $\mathbf{x}$ as $\mathbf{x}^{(m,W,O)}$, $m \in \mathcal{S}' = \{0, 1, \ldots, U'V'-1\}$, $U' = \left\lfloor \frac{U-O}{W-O} \right\rfloor$, $V' = \left\lfloor \frac{V-O}{W-O} \right\rfloor$, with the upper-left corner of the m-th window corresponding to site

$$i = i(m, W, O) = (W-O) \cdot \left(m + \left(U - U'\right) \cdot \left\lfloor \frac{m}{U'} \right\rfloor \right).$$

Set $\mathcal{N}_m$ then contains the four-connected neighbors with respect to site m, $\mathcal{N}_m = \{m-1, m+1, m-U', m+U'\}$, ignoring boundary conditions for the sake of simplicity. Windows are said to be fully overlapping, if $O = W-1$. Overlap $O = 0$ results in non-overlapping windows. With a slight abuse of notation, also note that $\mathbf{x}^{(m,1,0)} = x_m$. A *label field* $\mathbf{y}$ is the result of a mapping $\mathcal{X} \to \mathcal{Y} = \{-1, 1\}^{|\mathcal{S}'|}$ that indicates which (if any) sites are part of an image manipulation, i. e., an image region that is not genuinely the output of the image's source camera. Without loss of generality, we assume that negative labels indicate a manipulation. Finally, recall that the normalized correlation between two vectors $\mathbf{a}$ and $\mathbf{b}$ with sample means $\bar{\mathbf{a}}$ and $\bar{\mathbf{b}}$, respectively, is

$$\operatorname{corr}(\mathbf{a}, \mathbf{b}) = \frac{\langle \mathbf{a} - \bar{\mathbf{a}}, \mathbf{b} - \bar{\mathbf{b}} \rangle}{\|\mathbf{a} - \bar{\mathbf{a}}\| \|\mathbf{b} - \bar{\mathbf{b}}\|}.$$

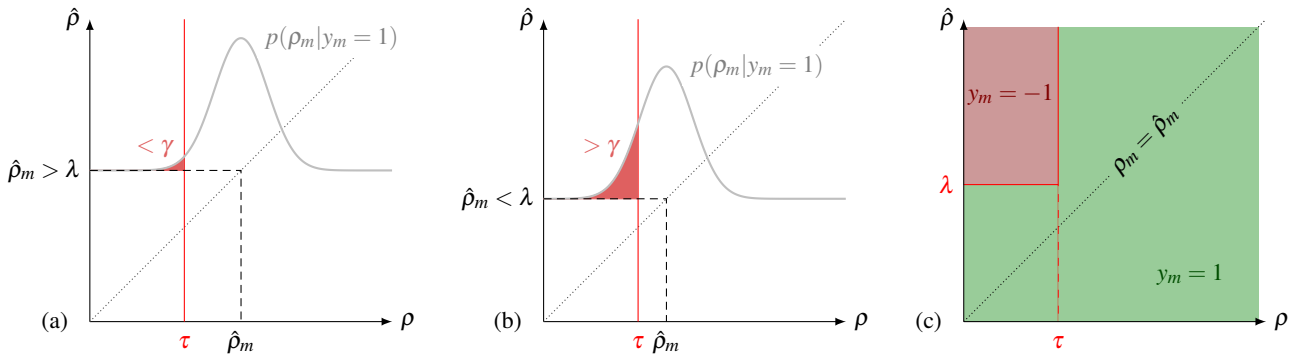

Figure 1: Interplay between correlation ρ and predicted correlation $\hat{\rho}$ in PRNU-based manipulation detection, where a manipulation is declared ($y_m = -1$) if $\rho_m < \tau$ and $\hat{\rho}_m > \lambda$. Threshold λ may be linked to a desired probability of false alarm γ, as obtained from modeling $p(\rho_m|y_m = 1)$. (a) A high predicted correlation $\hat{\rho}_m$ suggests a low probability of falsely labeling a genuine region with low correlation $\rho_m < \tau$ as manipulated. (b) Lower predicted correlations increase the risk of false positives. (c) Equivalent decision regions.

PRNU-based Image Forensics

The sensor fingerprint of a digital camera can be estimated from a sufficient number of genuine camera outputs $\mathbf{x}_1, \ldots, \mathbf{x}_L$, each of which is assumed to contain the camera-specific PRNU term. Adopting a multiplicative noise model $\mathbf{x} = (1+\mathbf{k}) \cdot \tilde{\mathbf{x}} + \boldsymbol{\theta}$, with clean sensor output $\tilde{\mathbf{x}}$ and additive i.i.d. Gaussian modeling noise $\boldsymbol{\theta}$, the maximum likelihood estimator of the $U \times V$ PRNU factor $\mathbf{k}$ is [2]

$$\hat{\mathbf{k}} = \left(\sum_{l=1}^{L} \mathbf{w}_l \mathbf{x}_l \right) \cdot \left(\sum_{l=1}^{L} \mathbf{x}_l^2 \right)^{-1}, \tag{1}$$

where $\mathbf{w}_l = \mathbf{x}_l - F(\mathbf{x}_l)$ is the noise residual obtained by feeding the l-th image into a suitable denoising filter $F(\cdot)$. A post-processing step helps cleaning estimate $\hat{\mathbf{k}}$ from so-called non-unique artifacts, e. g., due to demosaicing or lens distortion correction [2, 5].

With a pre-computed camera fingerprint estimate $\hat{\mathbf{k}}$, a query image $\mathbf{x}$ from the same camera can be analyzed for manipulations in a sliding-window manner by inspecting the normalized correlation $\boldsymbol{\rho} = (\rho_m)$, $m \in \mathcal{S}'$, between the noise residual from the m-th $W \times W$ window, $\mathbf{w}^{(m,W,O)} = (\mathbf{x} - F(\mathbf{x}))^{(m,W,O)}$, and the corresponding portion of the PRNU term, $(\mathbf{x}\hat{\mathbf{k}})^{(m,W,O)}$,

$$\rho_m = \operatorname{corr}\left(\mathbf{w}^{(m,W,O)}, (\mathbf{x}\hat{\mathbf{k}})^{(m,W,O)} \right). \tag{2}$$

In line with prior work, we assume fully overlapping windows ($O = W - 1$). In the most simple setting, the detector decides

$$y_m = \operatorname{sgn}(\rho_m - \tau) \tag{3}$$

for a suitably chosen threshold $\tau > 0$ [6]. In this context, a *true positive* refers to a local decision $y_m = -1$ when the corresponding analysis window is indeed part of a manipulation. On the contrary, a *false positive* will occur when a negative label is assigned to a genuine region. Larger thresholds τ generally increase the true positive rate at the expense of more false positives.

A major practical challenge with PRNU-based manipulation detection is that the strength of local correlation in Equation (2) can vary greatly with image content. Overly dark, saturated, or textured areas may yield low similarity scores even in the absence of a manipulation, partly due to the multiplicative nature of the PRNU noise, and partly due to imperfections of the denoising filter. A remedy is to include a correlation predictor $\hat{\boldsymbol{\rho}} = (\hat{\rho}_m)$, $m \in \mathcal{S}'$, that indicates the expected fingerprint strength by modeling $p(\rho_m|y_m = 1)$ as Generalized Gaussian with mean $\hat{\rho}_m$ [6]. A linear predictor based on local image characteristics $\phi_c\left(\mathbf{x}^{(m,W,O)}\right) \in \mathbb{R}$,

$$\hat{\rho}_m = \sum\nolimits_c \beta_c \cdot \phi_c\left(\mathbf{x}^{(m,W,O)}\right), \tag{4}$$

defined in terms of the quadratic expansion of three simple features, intensity, flatness, and texture, works sufficiently well [6]. The nine optimal, in a least squares sense, linear regression coefficients β_c can be determined from a small set of genuine images in advance. The rationale here is that more conservative decisions should be in place when the local correlation cannot be expected to take on large values *per se*. In other words, an adjusted label assignment rule decides $y_m = -1$ iff ρ_m falls below a certain threshold τ *and* the probability of making a false positive error is small [6],

$$\int_{-\infty}^{\tau} p(\rho_m|y_m = 1)\, \mathrm{d}\rho_m < \gamma. \tag{5}$$

Equivalently, this translates to a decision rule

$$y_m = \tfrac{1}{2}\big(1 - \operatorname{sgn}(\tau - \rho_m) \cdot \operatorname{sgn}(\hat{\rho}_m - \lambda) - \operatorname{sgn}(\tau - \rho_m) - \operatorname{sgn}(\hat{\rho}_m - \lambda)\big), \tag{6}$$

requesting that $\rho_m < \tau$ and $\hat{\rho}_m > \lambda$ [3], see also Figure 1.

The label assignment rules in Equations (3) and (6) reach decisions for each sliding window $\mathbf{x}^{(m,W,O)}$ independently, i. e., information about / from surrounding windows is not taken into account in this process. Chierchia et al. [3] deviate from this approach by formulating the problem as a *global* label mapping

$$\mathbf{y}^* = \arg\max_{\mathbf{y} \in \mathcal{Y}} p(\mathbf{y}|\mathbf{x}), \tag{7}$$

with a posterior refined to reflect that decisions should be made based on the local correlations $\boldsymbol{\rho}$ and the predicted correlations $\hat{\boldsymbol{\rho}}$,

$$\mathbf{y}^* = \arg\max_{\mathbf{y} \in \mathcal{Y}} p(\mathbf{y}|\boldsymbol{\rho}, \hat{\boldsymbol{\rho}}). \tag{8}$$

The solution $\mathbf{y}^*$ is a $U' \times V'$ matrix, in correspondence with $\mathcal{S}'$. Invoking Bayes' rule and acknowledging that the predicted correlation depends solely on the image content leads to

$$\mathbf{y}^* = \arg\max_{\mathbf{y}\in\mathcal{Y}} \; p(\boldsymbol{\rho}|\hat{\boldsymbol{\rho}},\mathbf{y}) \cdot p(\mathbf{y}) \,, \tag{9}$$

a formulation that conveniently makes prior assumptions about the label field $\mathbf{y}$ explicit. Chierchia et al. [3] propose an Ising model over single-site cliques and four-connected two-site cliques [7] to factor in that labels should possess a certain smoothness. In other words, the authors wish to enforce that labels change only gradually. The resulting prior is a Markov random field (MRF) that has a Gibbs distribution

$$p(\mathbf{y}) = \frac{1}{Z} \exp\left(-\frac{\alpha}{2} \sum_m y_m - \frac{\beta}{2} \sum_m \sum_{n\in\mathcal{N}_m} (1 - y_m y_n) \right) , \tag{10}$$

with $\alpha = \log(\Pr(y=-1)/\Pr(y=1))$ and edge penalty β. Z is a normalizing constant. With an additional conditional independence assumption, $p(\boldsymbol{\rho}|\hat{\boldsymbol{\rho}},\mathbf{y}) = \prod_m p(\rho_m|\hat{\rho}_m, y_m)$, the posterior may be rewritten as

$$p(\mathbf{y}|\boldsymbol{\rho},\hat{\boldsymbol{\rho}}) \propto \exp\left(\sum_m \log p(\rho_m|\hat{\rho}_m, y_m) - \frac{\alpha}{2} \sum_m y_m + \frac{\beta}{2} \sum_m \sum_{n\in\mathcal{N}_m} y_m y_n \right) . \tag{11}$$

We refer to [3] for a numerical solution to Equation (9), where a Gaussian model is adopted for the likelihood terms in Equation (11). Experimental results suggest that the global MRF formulation outperforms detectors with independent label assignments.

Discriminative Random Fields

Discriminative random fields (DRFs) [4] have been proposed as a flexible solution to probabilistic label mapping problems in the general form of Equation (7). Contrary to the Bayesian reformulation with an Markov random field (MRF) label prior, a DRF models the posterior $p(\mathbf{y}|\mathbf{x})$ directly as an MRF, without modeling the prior $p(\mathbf{y})$ and the likelihood $p(\mathbf{x}|\mathbf{y})$ individually. Hence, a DRF is technically a conditional random field [8], which can be defined formally as a probabilistic graphical model over a set of input variables $\mathbf{X} = (X_1, X_2, ..., X_N)$ and a set of output variables $\mathbf{Y} = (Y_1, Y_2, ..., Y_N)$ [9]. Variables are indexed by the vertices of a graph $\mathcal{G} = (\mathcal{S}, \mathcal{E})$, and they obey the Markov property with respect to the graph when conditioned on $\mathbf{X}$, i. e., $p(y_i|\mathbf{x}, \mathbf{y}_{\mathcal{S}-\{i\}}) = p(y_i|\mathbf{x}, \mathbf{y}_{\mathcal{N}_i})$, where $\mathcal{S} - \{i\}$ is the set of all nodes in G excluding site i, and $\mathcal{N}_i$ is the set of neighbors of site i. Assuming only up to pairwise clique potentials, Kumar and Hebert [4] propose the following general posterior model:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z} \exp\left(\sum_{i\in S} A_i(y_i, \mathbf{x}) + \sum_{i\in S} \sum_{j\in\mathcal{N}_i} I_{i,j}(y_i, y_j, \mathbf{x}) \right) \tag{12}$$

Quantities $-A_i(y_i, \mathbf{x})$ and $-I_{i,j}(y_i, y_j, \mathbf{x})$ are the association potentials and interaction potentials, respectively, which we assume to be homogenous and isotropic here for the sake of simplicity. These potentials are built upon discriminative models, giving the DRF its name. Association potentials can be thought of as local discriminators and reflect individual local decisions. Specifically, the association of site i with a certain label instance y_i, given a suitable r-dimensional local feature representation $\boldsymbol{\psi}(\mathbf{x}, i)$ of the observed data $\mathbf{x}$, is modeled after

$$A_i(y_i, \mathbf{x}) = \log p(y_i|\boldsymbol{\psi}(\mathbf{x}, i)) \,, \qquad \boldsymbol{\psi} : \mathcal{X} \times \mathcal{S} \to \mathbb{R}^r \,. \tag{13}$$

Interaction potentials act as *data-dependent* label smoothing function and may be modeled similar to the association potentials above as pairwise discriminative models,

$$I_{i,j}(y_i, y_j, \mathbf{x}) = \log p(y_i y_j|\boldsymbol{\zeta}(\mathbf{x}, i, j)) \,. \tag{14}$$

Function $\boldsymbol{\zeta} : \mathcal{X} \times \mathcal{S} \times \mathcal{S} \to \mathbb{R}^s$ extracts suitable features related to sites i and j from $\mathbf{x}$ that reflect how strongly labels y_i and y_j should concur, i. e., neighboring sites should be assigned the same label iff supported by data. The influence of observations at neighboring sites $j \in \mathcal{N}_i$ on the label at site i is thus made explicit. This is fundamentally different from MRF label field priors, which are generally independent of data. In addition, both potentials may in general depend on *all* observations. Conditional and discriminative random fields have been successfully applied to a variety of problems in the field of image segmentation, object recognition, and computer vision in general [10–14].

Design of Potentials

We continue with instantiating the general DRF formulation in Equation (12) for the specific problem of PRNU-based manipulation detection.

Association Potentials

Following prior art, we work with local correlations $\boldsymbol{\rho}$ and predicted correlations $\hat{\boldsymbol{\rho}}$. In the DRF framework, association potentials are then a local discriminative model to express how strongly site $m \in \mathcal{S}'$ favors label $y_m \in \{-1, 1\}$ given $\boldsymbol{\rho}$ and $\hat{\boldsymbol{\rho}}$, hence $\boldsymbol{\psi}(\mathbf{x}, m) = \boldsymbol{\psi}(\boldsymbol{\rho}, \hat{\boldsymbol{\rho}}, m)$. Instead of a straightforward translation of the setup from the previous section, which would consider $\boldsymbol{\psi}(\boldsymbol{\rho}, \hat{\boldsymbol{\rho}}, m) = (\rho_m, \hat{\rho}_m)$, we incorporate the two following conceptual changes.

First, we wish to model local decisions based on the *differences* $\boldsymbol{\Delta} = \hat{\boldsymbol{\rho}} - \boldsymbol{\rho}$, for which a threshold-based detector with threshold η would decide

$$y_m = \operatorname{sgn}(\eta - \Delta_m) \,. \tag{15}$$

Hence, large differences are likely to be associated with manipulated regions, see also Figure 2. The premise here is that we assume the correlation predictor to be an accurate and consistent model of the expected correlation in the absence of a manipulation. Therefore, the label assignment rule above may also label sites with relatively high correlation as manipulated, if the measured correlation ρ_m is considerably smaller than $\hat{\rho}_m$. At the same time, and more importantly, a comparison of Figures 1 (c) and 2 indicates that the adjusted rule is much more likely to label sites with low correlation as manipulated (depending on the specific threshold settings). Overall, Equation (15) thus puts more emphasis on the correlation predictor than Equation (6).

A second difference to before is that we allow the detector to work with *aggregates of local characteristics*. Specifically, after obtaining the difference field $\boldsymbol{\Delta}$ from fully overlapping $W \times W$

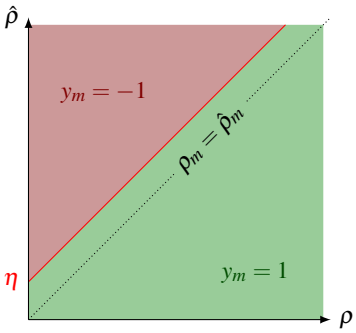

Figure 2: Adjusted decision regions based on correlation difference $\Delta_m = \hat{\rho}_m - \rho_m \gtrless \eta$.

analysis windows, we consider sites $\bar{m} \in \bar{\mathcal{S}}$ that result from averaging the correlation differences over small *non-overlapping* $B \times B$ blocks $\mathbf{\Delta}^{(\bar{m},B,0)}$,

$$\bar{\Delta}_{\bar{m}} = \frac{1}{B^2} \sum_{k=0}^{B^2-1} \Delta_k^{(\bar{m},B,0)}, \tag{16}$$

with $\bar{\Delta}_{\bar{m}} = \Delta_m$ for $B = 1$. Averaging attenuates the impact of measurement noise. In addition, working with non-overlapping blocks is effectively equivalent to sub-sampling, which can cut the computational load of inference routines substantially.

As for the label conditionals in Equation (13), we work with a logistic function model,

$$\exp(A_{\bar{m}}) = \frac{1}{1+\exp\left(-y_{\bar{m}}\left(\omega_0 + \omega_1 \bar{\Delta}_{\bar{m}}\right)\right)}, \tag{17}$$

thereby translating the hard threshold in Equation (15) to a probabilistic discriminative setting, $y_{\bar{m}} = \mathrm{sgn}(A_{\bar{m}} - 0.5)$. Parameters $\boldsymbol{\omega} = (\omega_0, \omega_1)$ of the above sigmoid have intuitive interpretations. The second parameter, $\omega_1 < 0$, determines the general shape of the sigmoid, with larger absolute values yielding sharper transitions between the two classes, see also Figure 3. For a fixed ω_1, parameter ω_0 then effectively controls the threshold η, as it follows directly from that $\eta = -\omega_0/\omega_1$.

Interaction Potentials

For the interaction potentials, we make the simple assumption that large absolute differences $|\bar{\Delta}_{\bar{m}} - \bar{\Delta}_{\bar{n}}| > \delta > 0$ between neighboring sites should impose different labels. We use a sigmoid model for pairwise equal labels to this end,

$$\exp(I_{\bar{m},\bar{n}}) = \frac{1}{1+\exp(-y_{\bar{m}} y_{\bar{n}}(\nu_0 + \nu_1 |\bar{\Delta}_{\bar{m}} - \bar{\Delta}_{\bar{n}}|))}, \tag{18}$$

which we evaluate for four-connected neighborhoods as described before. The interpretation of the model parameters $\boldsymbol{\nu} = (\nu_0, \nu_1)$ parallels the discussion on association potentials above. Specifically, threshold δ may be expressed as $\delta = -\nu_0/\nu_1$, and $\nu_1 < 0$.

Model Parameters and Inference

With the association and interaction potentials described above, and given a set of weights $(\boldsymbol{\omega}, \boldsymbol{\nu})$, our goal is ultimately to infer the optimal label assignments from the posterior model

$$p(\mathbf{y}|\bar{\mathbf{\Delta}}) \propto \exp\left(\sum_{\bar{m}} A_{\bar{m}} + \sum_{\bar{m}} \sum_{\bar{n} \in \mathcal{N}_{\bar{m}}} I_{\bar{m},\bar{n}} \right). \tag{19}$$

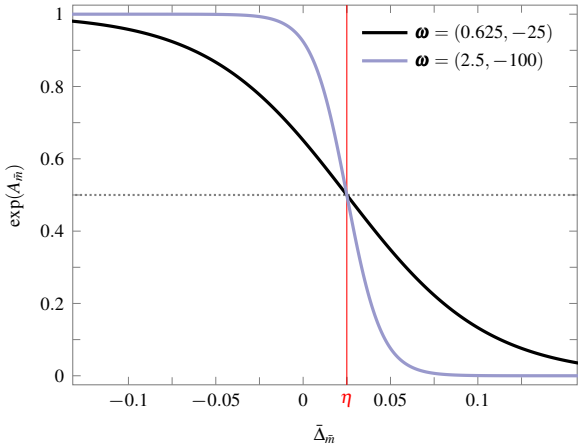

Figure 3: Sigmoid class conditional model for threshold $\eta = 0.025$ with two different weight settings.

Various standard routines are available for problems of this type [4, 14]. We chose to employ loopy belief propagation (LBP) to solve for the maximum a posteriori (MAP) marginals $y^*_{\bar{m}} = \arg\max_{y_{\bar{m}}} p(y_{\bar{m}}|\bar{\mathbf{\Delta}})$ as implemented in the CRF2D toolbox[1] for our exploratory experiments, acknowledging that other techniques may turn out to be more suitable, however [15].

While the model weights are typically learnt from a representative set of training data for optimal inference in a general DRF setting [4], we decided to omit this step as our setup entails an easy-to-interpret framework of association and interaction potentials. A grid search over suitable candidate settings $(\boldsymbol{\omega}, \boldsymbol{\nu})$ seems sufficient. Of particular interest are the influence of sharp vs. smooth potential transitions (see Figure 3) and indicative settings for the effective association and interaction thresholds η and δ, respectively.

It is finally worth pointing out that the MAP label assignment inferred from the above DRF formulation reduces to the threshold-based detector in Equation (15) when all interaction terms are set to a constant.

Experimental Setup

We work with a subset of 125 never-compressed Adobe Lightroom images from the Dresden Image Database [16], all taken by a Nikon D70 digital camera. All images were cropped to a common size of 801×801 pixels and converted to grayscale before any further processing. The camera's fingerprint was estimated from 25 homogeneously lit flat field images, applying the post-processing suggested in [2]. All noise residuals in our experiments were computed with the "standard" Wavelet denoising filter [17]. Image manipulations were simulated by randomly replacing a square region from the center of each test image with randomly selected content of equal size from another image taken by a different camera. We consider manipulations of size 384×384 and 128×128. Images were analyzed with fully overlapping analysis windows of sizes 128×128 and 64×64, respectively. The correlation predictor in Equation (4) was trained for each window size W on 4485 blocks from 13 clean images. We report true positive and false positive rates on pixel level, averaged over all test images. Different parameter and threshold settings yield different discrete operating points. Baseline results from the plain threshold-based

[1] `http://www.cs.ubc.ca/~murphyk/Software/CRF/crf2D_usage.html`

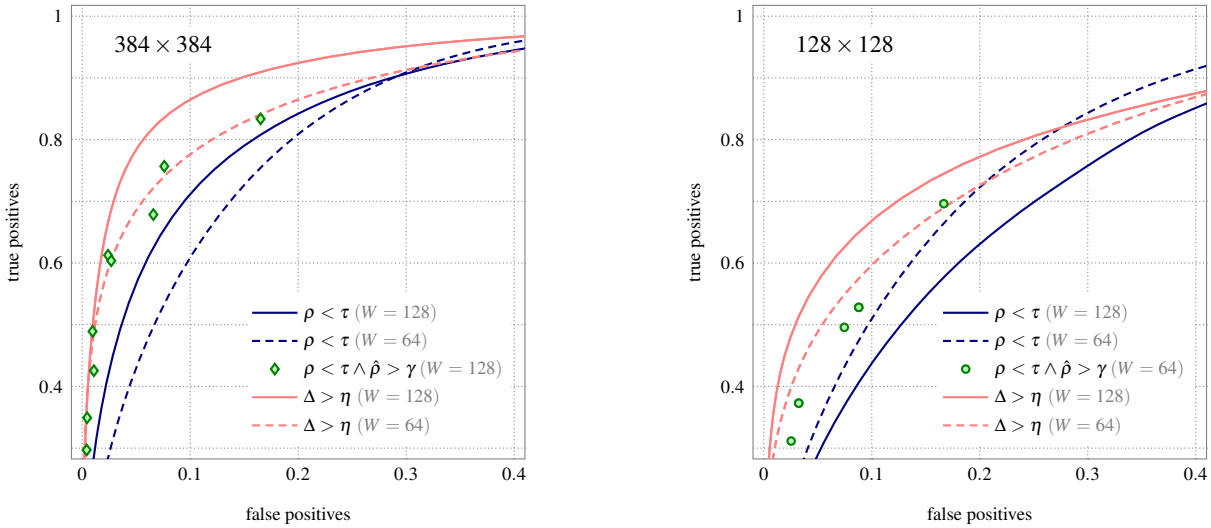

Figure 4: Baseline localization performance of various threshold-based detectors with independent label assignments for manipulations of size 384×384 (left) and 128×128 (right). Detectors: (3) blue; (6) green; (15) red. Indicative operating points for detector (6) were obtained from select well-performing threshold pairs (τ, γ).

detectors in Equations (3) and (15) will be presented in the form of ROC curves. Label maps obtained from aggregated correlation differences were resized by a factor of B to restore the original image size. Other than that, all results reported in this paper are based on "plain" label maps, without any further post-processing (such as dilation) for the sake of better cross-technique comparability.

Results

Figure 4 sets the baseline for our experimental results. The graphs report the localization performances of various PRNU-based detectors that reach their decision on a window-by-window basis, without explicitly taking information from surrounding windows into account. The left panel presents results for manipulations of size 384×384. Corresponding results for the smaller 128×128 manipulations are depicted on the right. Specifically, we include ROC curves for thresholding the correlation as in Equation (3) and the difference between predicted and measured correlation as in Equation (15), computed from windows of size 128×128 and 64×64, respectively. Indicative operating points for the detector in Equation (6) with a window size of 128×128 are presented for a number of well-performing threshold pairs (τ, γ). The smaller window size ($W = 64$) did not achieve competitive results in our experiments, so we omit it here.

All detectors expectedly perform better when a large region lacks the genuine sensor noise fingerprint. More interestingly, observe that the proposed correlation differences Δ_m computed over 128×128 windows yield the best results in the most relevant scenario where low false positive rates are concerned. A comparison of Figures 1 (c) and 2 suggests that the increased true positive rates can be attributed to the different treatments of low-correlation regions. All following results will thus include ROC curves obtained from thresholding $\Delta_m \gtrless \eta$ with $W = 128$ as a reference.

Moving on to detectors that make decisions by taking label interdependencies into account, Figure 5 reports results obtained from global MAP label assignments based on the MRF posterior in Equation (11) and the DRF posterior in Equation (19), respectively. As before, the results for large and small manipulations are split between the left and the right panel. Operating points of the DRF detector ($B = 2$) were obtained by setting the association potential model weights (ω_0, ω_1) to resemble a select set of five different effective decision thresholds $\eta = -\omega_0/\omega_1$, cf. Equation (15). We observed that a sharp transition profile with $\omega_1 = -100$ (see Figure 3) gave preferable results in most situations, except for small-sized manipulations analyzed with small windows ($W = 64$), where a smoother transition with $\omega_1 = -25$ was more suitable. This seems intuitive, as this allows the association potentials more freedom in their label assignments in the presence of less reliable decisions from smaller analysis windows, while the interaction potentials will achieve label smoothness relatively easily when large manipulated regions are concerned. The effective interaction threshold was set to $\delta = 0.1$ in combination with a sharp potential transition in all depicted settings. Figure 5 suggests that the explicit consideration of label interdependencies in the DRF formulation affects pixel-level false positive and true positive rates positively when relatively large manipulations are concerned. For small-sized manipulations, the effect is less pronounced and only measurable for higher false positive rates. Overall, our experiments seem to point to the conclusion that it is first and foremost the adjusted discriminative model based on the correlation differences Δ_m that leads to considerable improvements over the various depicted MRF operating points (each corresponding to a unique parameter tuple (α, β) with $\exp(\alpha) \in \{0.25, 0.5, 0.75\}$ and $0.03 \leq \beta \leq 1000$). At the same time however, also observe that the DRF *does* yields a substantial performance gain over independent label assignments from thresholding Δ_m when small windows ($W = 64$) are to be considered. This effect is most prominent for larger manipulations, where the small analysis windows even outperform the bigger ones.

Figure 6 emphasizes that the aggregation of correlation differences over small non-overlapping $B \times B$ blocks is integral to the strong performance of small analysis windows. The graphs indicate that switching from $B = 1$ (i. e., no aggregation) to $B = 2$ can result in an immense reduction of false positives, in particu-

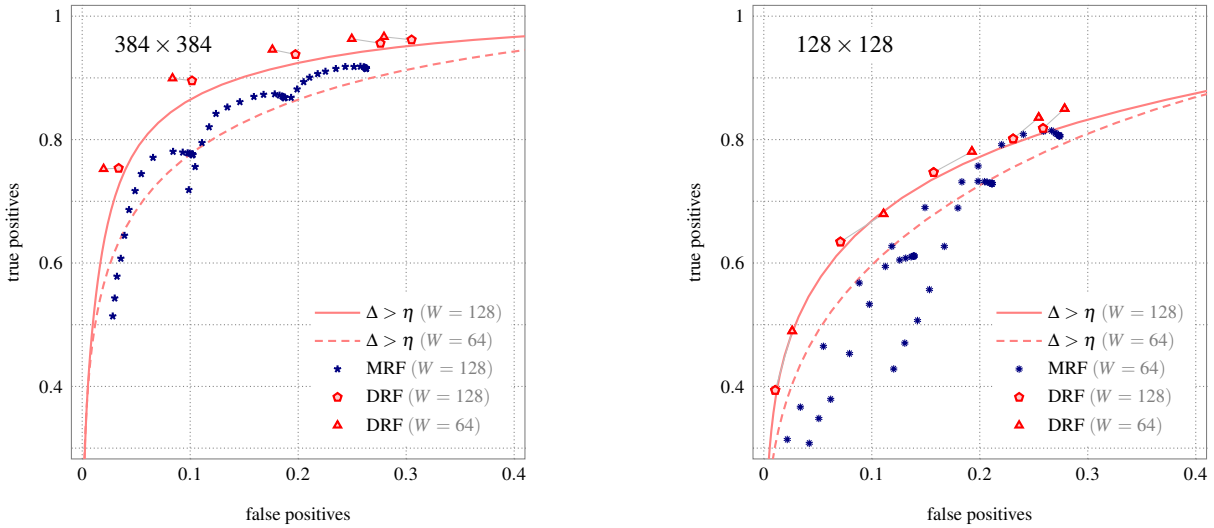

Figure 5: Localization performance of global label mapping approaches for manipulations of size 384×384 (left) and 128×128 (right). Proposed DRF with window sizes $W \in \{64, 128\}$; aggregation over non-overlapping 2×2 blocks ($B = 2$). Indicative DRF operating points were obtained for a set of effective association thresholds with a fixed effective interaction threshold δ. A gray line connecting a pair of operating points from different window sizes indicates that the detectors operated with the same effective association threshold. MRF operating points were obtained for different parameter pairs (α, β), $\exp(\alpha) \in \{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}$.

lar for manipulations of larger size (all other parameter settings parallel the setup in Figure 5). Increasing the aggregation block size further does generally not boost performance more, however. Also note that ROC curves obtained from thresholding aggregated correlation differences are largely equivalent to the ones obtained with $B = 1$, which are depicted in all of our graphs.

Throughout our experiments, we observed that large effective interaction thresholds δ are needed for reliable pixel-level manipulation localization. We refer to Figure 7 for a comparison of operating points obtained with thresholds $\delta = 0.1$ and $\delta = 0.01$, keeping all other parameter settings unchanged. In general, we found that larger thresholds gave smoother maps, in line with the design of the interaction potential function. Considering that the values of the Δ_m's in genuine and non-genuine regions are on average only about 0.06 apart, a few notes are in order regarding the relatively large values of δ. As fully overlapping analysis windows generally impose relatively gradual changes in the difference field $\mathbf{\Delta}$, sudden extreme changes are unlikely, also at the edges between genuine and manipulated areas. Hence, it is extremely unlikely to encounter sites where $|\Delta_m - \Delta_n| > 0.1$. A large threshold, in particular in combination with a sharp transition profile as the one favored in our experiments, thus effectively forces the label field to be continuous, relatively independently of the observed data. We found that increasing the effective interaction threshold will typically reduce the false positives, against the backdrop of also decreasing the true positive rate as δ gets larger. In other words, the inference routine will "shrink" the labeled manipulated region to an increasingly smaller area where a reliable decision can be made. The optimal δ depends on the association potential threshold. For operating points under a low false alarm regime, a lower interaction threshold can be more beneficial. Relatively smaller effective interaction thresholds are also preferable when the analysis window size approaches the size of the manipulated region. It is worth mentioning here that we experienced large effective interaction thresholds to cause convergence issues for the loopy belief propagation algorithm as the pairwise potentials in Equation (18) approach 0 and 1 numerically. Working with non-overlapping aggregator blocks helped to attenuate this problem greatly. Overall, the preference of interaction potentials that are quasi-independent of the observed data deserves further exploration, as it may hint to a more general problem that could be inherent to any global label mapping problem when label interaction is made dependent on differences between features computed from fully overlapping analysis windows.

Figure 8 closes this section and presents a number of illustrative label maps obtained with the MRF and DRF detectors for window size $W = 64$ (image size 1000×1000 pixels). Parameter settings reflect the discussion of Figure 5 above. Specifically, the MRF detector operated with $\alpha = \log(0.25)$ and $\beta = 250$. The effective DRF thresholds were set to $\eta = 0.02$ (with $\omega_1 = -25$) and $\delta = 0.1$ (with $\nu_1 = -150$), corresponding to the operating point with lowest false positive rate in the left panel of Figure 5.

Concluding Remarks

We have explored a discriminative random field (DRF) formulation to frame PRNU-based image manipulation localization as a probabilistic binary labeling problem. Assuming that the source camera of the query image is known, our detector operates in a sliding-window mode to assess local neighborhoods for the presence of the camera's sensor fingerprint by means of correlation. Association potentials in the form of local discriminators distinguish genuine from manipulated content based on the deviation of the measured correlation from the expected local correlation as estimated by a correlation predictor. Interaction potentials explicitly model pairwise dependencies between local decisions. While our experimental results suggest improvements over prior art, we see room for further advances in our future work, in particular with respect to small-sized manipulations.

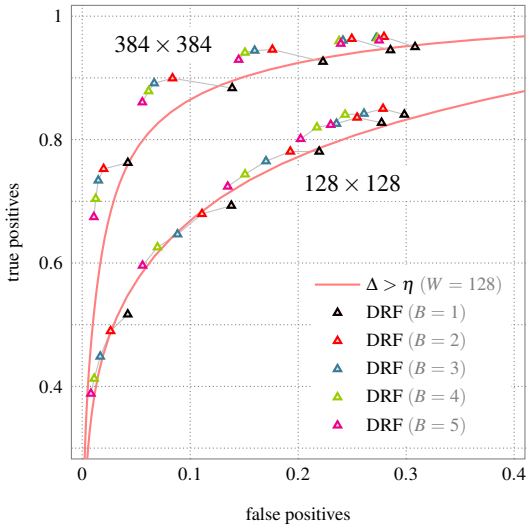

Figure 6: DRF performance ($W = 64$, $\gamma = 0.1$) for manipulations of size 384×384 and 128×128 with aggregations over $B \times B$ blocks, $B \in \{1, \ldots, 5\}$. Indicative operating points for a set of effective association thresholds. Gray lines connect groups of operating points with equal effective association thresholds and different B settings.

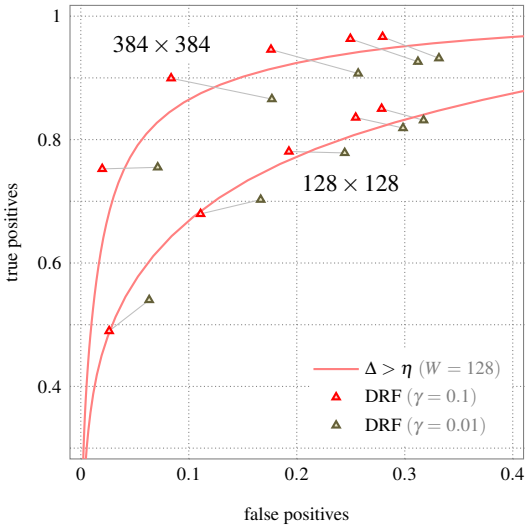

Figure 7: DRF performance ($W = 64$, $B = 2$) for manipulations of size 384×384 and 128×128 with effective interaction thresholds $\gamma \in \{0.01, 0.1\}$. Indicative operating points for a set of effective association thresholds. Gray lines connect pairs of operating points with equal effective association thresholds and different γ settings.

Specifically, we will consider alternative decision regions in the $(\rho, \hat{\rho})$-plane (cf. Figures 1 and 2) that do not penalize large local correlation values. An adjusted decision rule may label the region around site m as manipulated iff $\Delta_m > \eta$ *and* $\rho_m < \tau$. In a similar fashion, a second correlation threshold may introduce a small "safety margin" for very low correlation values to prevent false alarms. In addition, as image manipulations will often align with object boundaries, it seems viable to make the interaction potentials depend on image content. Guided filtering [18] and, very recently, image segmentation [19] are strategies that have been discussed in the literature before. This stream of research appears specifically relevant as our current choice of interaction potentials leads to label interactions that seem largely independent of the measured correlation quantities.

Overall, it remains to be seen to what degree small-sized image manipulations are detectable by means of PRNU-based techniques in general. As smaller manipulations require smaller analysis windows, any detector is bound by inherent limitations imposed by computing local correlation metrics over a relatively small number of samples. More critical benchmarks should thus not only examine more realistic image manipulations, but also consider alternative data-driven contenders [20, 21]

Acknowledgments

This material is based on research sponsored by DARPA and Air Force Research Laboratory (AFRL) under agreement number FA8750-16-2-0173. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA and Air Force Research Laboratory (AFRL) or the U.S. Government.

References

[1] R. Böhme and M. Kirchner, "Media forensics," in *Information Hiding*, S. Katzenbeisser and F. Petitcolas, Eds. Artech House, 2016, ch. 9, pp. 231–259.

[2] J. Fridrich, "Sensor defects in digital image forensics," in *Digital Image Forensics: There is More to a Picture Than Meets the Eye*, H. T. Sencar and N. Memon, Eds. Springer, 2013, pp. 179–218.

[3] G. Chierchia, G. Poggi, C. Sansone, and L. Verdoliva, "A Bayesian-MRF approach for PRNU-based image forgery detection," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 4, pp. 554–567, 2014.

[4] S. Kumar and M. Hebert, "Discriminative random fields," *International Journal of Computer Vision*, vol. 68, no. 2, pp. 179–201, 2006.

[5] T. Gloe, S. Pfennig, and M. Kirchner, "Unexpected artefacts in PRNU-based camera identification: A 'Dresden Image Database' case-study," in *ACM Multimedia and Security Workshop (MM&Sec)*, 2012, pp. 109–114.

[6] M. Chen, J. Fridrich, M. Goljan, and J. Lukáš, "Determining image origin and integrity using sensor noise," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 1, pp. 74–90, 2008.

[7] S. Li, *Markov Random Field Modeling in Image Analysis*. London: Springer-Verlag, 2009.

[8] C. Sutton and A. McCallum, "An introduction to conditional random fields," *Foundations and Trends in Machine Learning*, vol. 4, no. 4, pp. 267–373, 2011.

[9] D. Koller and N. Friedman, *Probabilistic Graphical Models*. MIT Press, 2009.

[10] S. Kumar and M. Hebert, "Discriminative fields for modeling spatial dependencies in natural images," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. Saul, and B. Schölkopf, Eds., 2004, pp. 1531–1538.

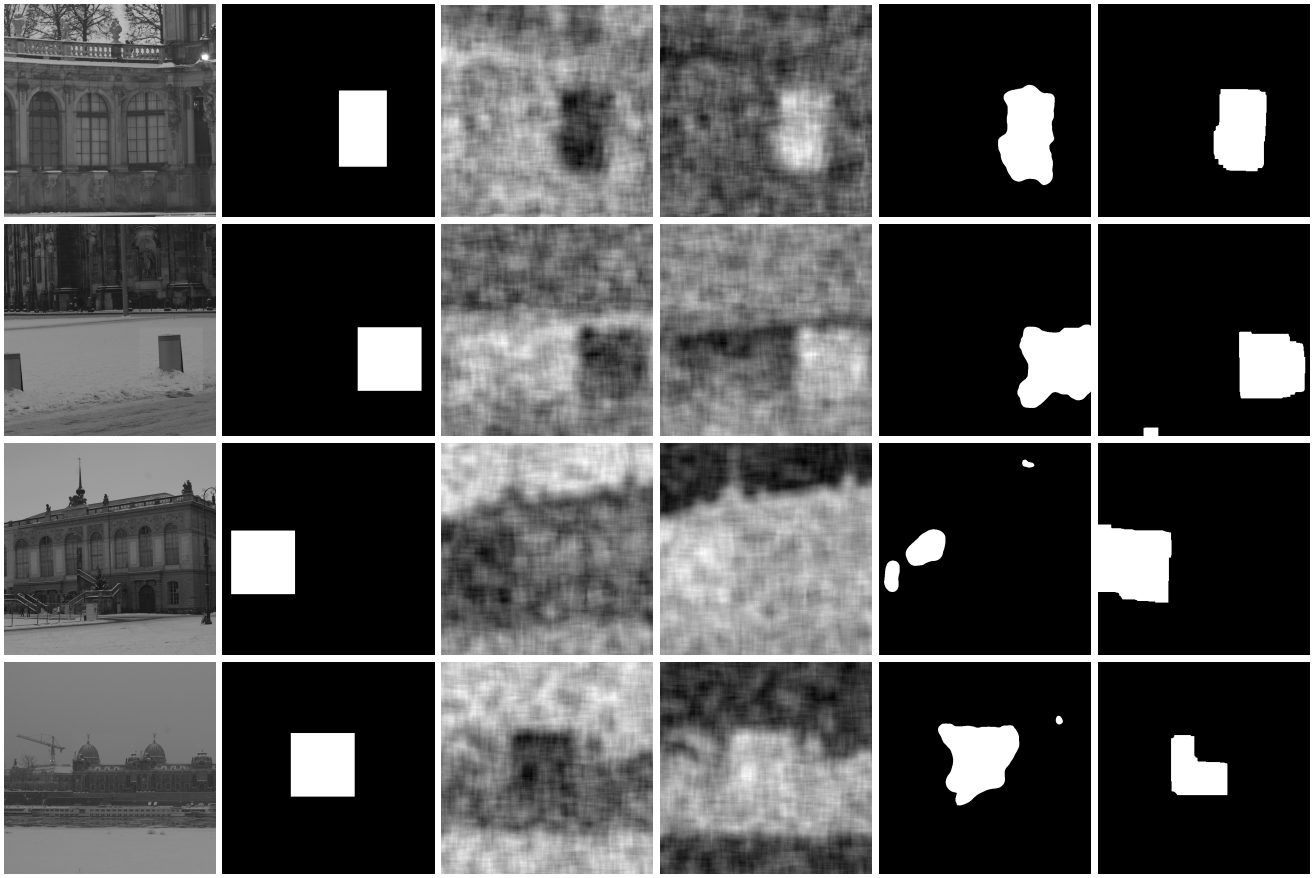

Figure 8: Manipulation localization with MRF and DRF detectors. From left to right: manipulated image; ground truth; local correlation ($W = 64$); difference between predicted and measured correlation; MRF label map; DRF label map. Image size 1000×1000 pixels.

[11] J. Verbeek and B. Triggs, "Scene segmentation with conditional random fields learned from partially labeled images," in *Advances in Neural Information Processing Systems 20*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds., 2008, pp. 1553–1560.

[12] N. Plath, M. Toussaint, and S. Nakajima, "Multi-class image segmentation using conditional random fields and global classification," in *International Conference on Machine Learning*, 2009, pp. 817–824.

[13] G. Roig, X. Boix, R. de Nijs, S. Ramos, K. Kühnlenz, and L. van Gool, "Active MAP inference in CRFs for efficient semantic segmentation," in *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 2312–2319.

[14] A. Blake, P. Kohli, and C. Rother, Eds., *Markov Random Fields for Vision and Image Processing*. MIT Press, 2011.

[15] K. Murphy, Y. Weiss, and M. I. Jordan, "Loopy belief propagation for approximate inference: An empirical study," in *Conference on Uncertainty in Artificial Intelligence (UAI)*, 1999, pp. 467–475.

[16] T. Gloe and R. Böhme, "The Dresden Image Database for benchmarking digital image forensics," *Journal of Digital Forensic Practice*, vol. 3, no. 2–4, pp. 150–159, 2010.

[17] 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*, vol. 6, no. 12, pp. 300–303, 1999.

[18] G. Chierchia, D. Cozzolino, G. Poggi, C. Sansone, and L. Verdoliva, "Guided filtering for PRNU-based localization of small-size image forgeries," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 6272–6276.

[19] P. Korus and J. Huang, "Multi-scale analysis strategies in PRNU-based tampering localization," *IEEE Transactions on Information Forensics and Security*, in press.

[20] D. Cozzolino, G. Poggi, and L. Verdoliva, "Splicebuster: A new blind image splicing detector," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2015.

[21] W. Fan, K. Wang, and F. Cayre, "General-purpose image forensics using patch likelihood under image statistical models," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2015.