IMAGE TAMPERING IDENTIFICATION USING BLIND DECONVOLUTION

Ashwin Swaminathan, Min Wu and K. J. Ray Liu

Electrical and Computer Engineering Department, University of Maryland, College Park

ABSTRACT

Digital images have been used in growing number of applications from law enforcement and surveillance, to medical diagnosis and consumer photography. With such widespread popularity and the presence of low-cost image editing softwares, the integrity of image content can no longer be taken for granted. In this paper, we propose a novel technique based on blind deconvolution to verify image authenticity. We consider the direct output images of a camera as authentic, and introduce algorithms to detect further processing such as tampering applied to the image. Our proposed method is based on the observation that many tampering operations can be approximated as a combination of linear and non-linear components. We model the linear part of the tampering process as a filter, and obtain its coefficients using blind deconvolution. These estimated coefficients are then used to identify possible manipulations. We demonstrate the effectiveness of the proposed image authentication technique and compare our results with existing works.

Index Terms – Multimedia forensics, image authentication, tampering detection.

1. INTRODUCTION

Digital images have been used in a growing number of applications from law enforcement, military, and reconnaissance to medical diagnosis and consumer photography. Such widespread popularity and the advent of low-cost sophisticated image editing softwares has led to a growing need for methods to ensure image integrity. Techniques such as semi-fragile image watermarking and robust image hashing have been proposed to establish the authenticity of the data [1]. However, these methods require that a signature be inserted at the time of creation of multimedia data. This impose several restrictions on its usage as many digital cameras and video recorders in the market still do not have the capabilities to add a watermark at the time of image creation. Hence, there is a strong motivation as a part of the emerging field of image forensics to devise *non-intrusive* methods to identify tampered images.

Most work on tampering detection literature identify tampering by studying the properties of a manipulated image in terms of the distortions it undergoes, which might include resampling [2], JPEG compression [3], lens distortions, Gamma correction, and additive noise [4]. Each of these processing operations modify the image statistics in a specific manner and thus can be identified by extracting certain salient features that would help distinguish such tampering from authentic data. For instance, when the image is upsampled, some of the pixel values are directly obtained from the smaller version of the image and the remaining pixels are interpolated and thus highly correlated with its neighbors [2]. Such post-processing operations involving interpolation can be identified by studying the induced correlations and solving for color interpolation coefficients [5]. Image manipulations such as contrast changes, Gamma correction and other image non-linearities have been modelled and higher order statistics such as the bispectrum have been used to identify them [4]. JPEG compression has been considered as quantization in the DCT domain and statistical analysis based on binning techniques have been used to estimate the quantization matrices [3].

Although these methods can be employed to identify the type and the parameters of the post-processing operation, it would require an exhaustive search over all the numerous kinds of post-processing operations to detect tampering. Classifier based approaches to detect tampering were proposed in [6, 11]. However, these methods require samples of tampered images to train the classifier for differentiating manipulated images from genuine ones. Further, these methods may not be able to efficiently identify tampering operations that are not modelled or considered directly as part of training. Thus, there is a strong need for a universal framework to distinguish authentic pictures from tampered images.

In this paper, we propose a novel technique based on blind deconvolution to verify the authenticity of a digital image. We consider the direct output images of a camera as untampered, and characterize its properties by a *ground-truth* imaging model. We assume that any further processing on the camera output can be represented as a combination of linear and non-linear components. We model the linear part as a *tamper filter* and find its coefficients using blind deconvolution. These estimated coefficients of the tampering block are compared to the delta function corresponding to no tampering, and a high similarity indicates that the test image is a direct output from a camera and is not manipulated. The proposed algorithm does not require any prior knowledge of the nature of the tampering operation, and can identify previously unseen manipulations.

2. SYSTEM MODEL

The proposed system model is shown in Fig. 1. We consider a genuine photograph as an output of the digital cameras' imaging process (point A in Fig. 1) and model any kind of further processing applied to it as a tampering block.

Email contact: {ashwins, minwu, kjrliu} @eng.umd.edu.

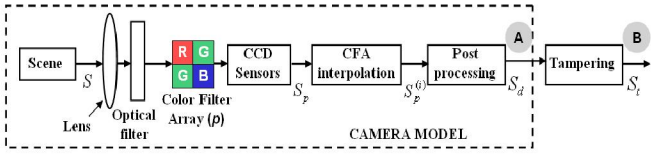

Fig. 1. System model for tampering detection

The details of the *ground-truth* imaging model are shown enclosed in dotted lines. The light from the scene pass through the lens and the optical filters and are finally recorded by the charge coupled device (CCD) detectors. Most digital cameras use a color filter array (CFA) to sample the real-world scene. The CFA consists of an array of color sensors, each of which captures the corresponding color of the real-world scene at an appropriate pixel location. To facilitate discussions, let S be the real-world scene to be captured by the camera and let p be the CFA pattern matrix. The CFA sampling converts the real-world scene S into S_p satisfying

$$S_p(x,y,c) = \begin{cases} S(x,y,c) & \text{if } p(x,y) = c, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

After the data obtained from the CFA is recorded, the intermediate pixel values (corresponding to the points where $S_p(x,y,c) = 0$ in (1)) are interpolated using its neighboring pixel values to obtain $S_p^{(i)}$. The color interpolation algorithm is typically proprietary to the particular camera manufacturer, and most companies may employ a different algorithm [7]. After interpolation, white balancing and color correction are done to remove unrealistic color casts [8]. Finally, the image may be JPEG compressed to reduce storage space to produce the output image S_d.

In our recent work [7], we propose component forensics as a methodology to reverse-engineer the imaging process and robustly estimate the algorithms employed in various components inside the digital camera. We show that the parameters of such important camera components as CFA and color interpolation can be non-intrusively estimated solely using output images. Our algorithm [7] estimates the color interpolation coefficients through texture classification and linear approximation, and finds the CFA pattern that minimizes the interpolation errors. Specifically, the image is divided into three types of regions based on the gradient features in a local neighborhood. Denoting $I_{x,y} = S_d(x,y,p(x,y))$. The horizontal and vertical gradients at the location (x,y) are found using

$$H_{x,y} = |I_{x,y-2} + I_{x,y+2} - 2I_{x,y}|, \quad (2)$$

$$V_{x,y} = |I_{x-2,y} + I_{x+2,y} - 2I_{x,y}|. \quad (3)$$

The image pixel at location (x,y) is classified into one of the three categories: *Region* $\Re_1$ contains those parts of the image with a significant horizontal edge for which $(H_{x,y} - V_{x,y}) > T$ where T is a suitably chosen threshold; *Region* $\Re_2$ contains those parts with $(V_{x,y} - H_{x,y}) > T$; and *Region* $\Re_3$ includes the remaining smooth parts of the image. Using the final camera output S_d, we obtain a set of linear equations for all the pixels in each region $\Re_i (i = 1,2,3)$, solving which we obtain the interpolation coefficients $\alpha_{\Re_i}$. Once these coefficients are estimated, they are used to re-interpolate the image and find the interpolation error. The CFA pattern that gives the lowest error gives the estimate of the CFA pattern.

3. PROPOSED TAMPERING DETECTION ALGORITHM

In this work, we build upon component forensics to develop robust image authentication systems for *verifying* if a given digital image is a direct camera output or not. Using the test image, we construct a *ground-truth* imaging model as described in Section 2, and estimate the model parameters such as CFA and color interpolation coefficients by employing component forensic methodologies [7][1]. These model parameters are used to estimate the camera output S_d and to find the coefficients of the tampering block. A high degree of similarity of the estimated tamper filter coefficients with the delta function indicates that the test image is an output of the camera and is therefore not tampered.

The tampering block coefficients can be found by separately minimizing the cost function J^c in each color channel

$$J^c(u) = \sum_{x,y} (\hat{S}_{te}(x,y,c) - S_{te}(x,y,c))^2 + \eta (\sum_{a,b} u(a,b,c) - 1)^2, \quad (4)$$

where S_{te} is the estimate of the camera output, and is obtained by filtering the corresponding color of the test input S_t with the coefficients of the inverse filter u. $\hat{S}_{te}$ is found from S_{te} by imposing the camera constraints given by

$$\hat{S}_{te}(x,y,c) = \begin{cases} \sum_{m,n} \alpha_{\Re_i}(m,n,c) S_{te}(x-m, y-n, c) \\ \qquad \forall \{x,y\} \in \Re_i, \text{ and } 1 \le c \le 3, \\ S_{te}(x,y,c) \qquad \text{otherwise.} \end{cases} \quad (5)$$

Here, in general, we may assume that $\sum_{m,n} u(m,n,c) = K$ for $c = 1,2,3$, where K is a constant. A value of $K = 1$ would ensure that the original image and its tampered version have similar brightness levels. Therefore, the cost function J^c in the c^{th} color component aims at minimizing the overall interpolation error and the deviation of the filter coefficient sum from 1. The value of η is used to adjust the weights of the relative individual costs.

The minimization problem can be solved using an recursive procedure as shown in Fig. 2. In the k^{th} iteration, we obtain our estimate of the original image $S_{te}^{(k)}$ by passing the test image S_t though the estimate of the inverse blurring filter $u^{(k)}$. We then impose the camera constraints as in (5) to obtain $\hat{S}_{te}^{(k)}$ and find the interpolation error. The inverse filter coefficients are then updated by [9]

[1]The camera model parameters obtained from the test image would be close to the actual parameters because the estimation algorithm is robust to moderate levels of post-processing operations [7].

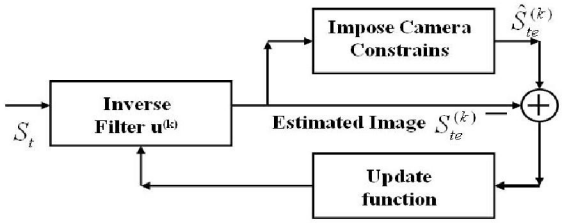

Fig. 2. A recursive algorithm to estimate the coefficients of the tampering block

$$u^{(k+1)} = u^{(k)} + t_k d_k, \quad \text{where} \tag{6}$$

$$d_k = \begin{cases} -\nabla J(u^{(k)}) & \text{if k = 0,} \\ -\nabla J(u^{(k)}) + \beta_{k-1} d_{k-1} & \text{otherwise,} \end{cases} \tag{7}$$

$$\beta_{k-1} = \frac{< \nabla J(u^{(k)}) - \nabla J(u^{(k-1)}), \nabla J(u^{(k)}) >}{||\nabla J(u^{(k-1)})||^2}, \tag{8}$$

and $J = \sum_{c=1}^{3} J^c$. The step sizes t_k are chosen by a line minimization algorithm that minimizes $J(u^{(k)} + t_k d_k) \leq J(u^{(k)} + t d_k)$. It can be shown, both in theory and simulations, that the optimization problem is convex and converges to a unique solution.

An alternative way to obtain the coefficients of the tampering block directly is by iteratively applying known constraints in both the pixel domain and in the Fourier domain [10]. The pixel domain constraints include the *camera* constraints as given in (5) and *boundedness* constraints restricting its values to the range $[0, 255]$. The Fourier domain constraints involves computing spectral response of the tamper filter based on the test image and the estimated camera output.

4. SIMULATION RESULTS AND DISCUSSIONS

We test the performance of our proposed framework with 100 images. We collect 25 different images from each of the 4 different cameras: (1) Canon A75, (2) Fujifilm S3000, (3) Sony P72, and (4) Minolta DiMage S304. These images are captured under completely random conditions – different scenaries, different lighting conditions, and compressed under different JPEG quality factors as specified by default values used in the camera. These images form our *camera data set*. These images were then processed to generate 27 different tampered versions per image by (1) resampling with percentage 50−150%, (2) JPEG compressing with quality factors 30−95, (3) adding noise of PSNR 5, 10 dB, (4) rotating with degrees 1−20, (5) average filtering with filter orders 3−11, and (6) median filtering with filter orders 3−7. These 2700 manipulated versions form the tampered image set.

The proposed blind deconvolution framework is implemented on all the images and the coefficients of the tampering block are computed in each case. In Fig. 3, we show the variation of the cost function J for the green color as a function of the number of iterations. We notice that the cost function converges in 10 iterations. Fig. 4 shows the frequency response of the estimated coefficients for an authentic image,

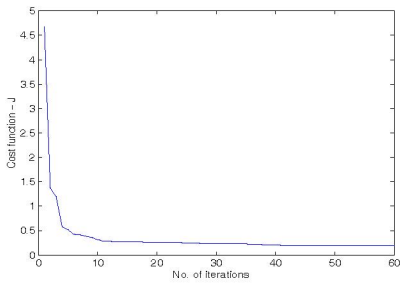

Fig. 3. Convergence of the cost function

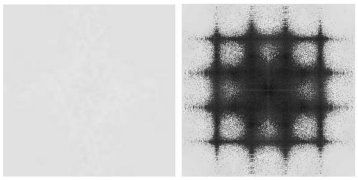

Fig. 4. Fourier transform of the tamper filter coefficients for (a) camera output (b) image spatially averaged with a 5×5 averaging filter. The frequency response is shown in the log scale and appropriately scaled for display.

and an image spatially averaged with a 5×5 averaging filter. We observe that the spectral response of the manipulation filter is almost a constant for an untampered image, and the corresponding spectrum for the filtered image shows distinctive periodic nulls as expected. These results indicate that the blind deconvolution algorithm performs well and is able to estimate the coefficients of the post-processing block with a reasonable accuracy.

Similarity Metric and Threshold Based Classifier: We design a *threshold based classifier* to distinguish manipulated images from authentic ones. Given the test input S_t, we find the frequency domain coefficients of the manipulation filter H_t, and compare it with the spectral response H_r obtained from a reference authentic camera output to measure the similarity among the coefficients. More specifically, we first find $LH_t = \log(H_t)$ and re-scale it to a $[0, 1]$ range to obtain the normalized logarithm of the frequency response (call it Θ_t). The similarity between the coefficients of the test input and the reference image is then found by comparing the corresponding normalized values

$$d(\Theta_t, \Theta_r) = \sum_{m,n} |\Theta_t(m,n) - \mu_t| \times |\Theta_r(m,n) - \mu_r| \tag{9}$$

where μ_t denotes the mean of the Θ_t and so on. A similarity value greater than a chosen threshold indicates that the image satisfies the ground-truth camera model and is authentic.

In Fig. 5(a) and (b), we show the average similarity scores found by comparing the coefficients obtained from tampered images (filtered and JPEG compressed) with the coefficients of the reference pattern (shown in Fig. 4(a)). We notice that

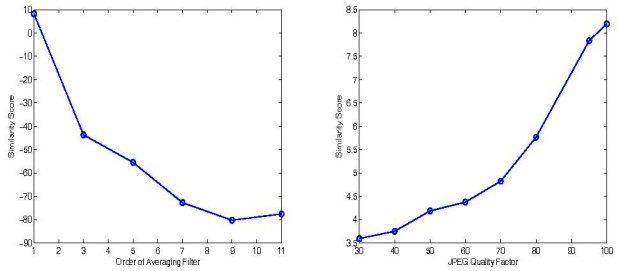

Fig. 5. Similarity scores between coefficients of the tampered image and the untampered reference image for (a) spatial averaging, (b) JPEG compression

the similarity scores reduce as the magnitude of the distortion increases. A similar trend is also observed for other kinds of manipulations.

Comparison Study: We compare our proposed tampering detection algorithm with the method in [11]. Here, the authors create a statistics vector by first extracting higher order moments from multiple level wavelet decompositions of the image. A linear predictor is then used to capture correlations that exist across orientation, space, and scale. An additional set of features is computed from the prediction error and the combined statistics vector is used in classification [11].

In our experiments using [11], we calculate the statistics vector for the entire set of 100 untampered and 2700 tampered images. A support vector machine (SVM) with a radial basis function (RBF) kernel is used for classification. We use a randomly chosen 50 untampered images along with its corresponding manipulated versions for training, and test on the remaining images. The fraction of correctly classified tampered images P_D, and the percentage of authentic images wrongly classified as tampered P_F are computed to obtain the receiver operating characteristics (ROC).

Fig. 6 shows the ROC for the scheme in [11] averaged over 100 iterations. The corresponding ROC curve for the proposed scheme obtained using the threshold based classifier is also shown alongside for comparison. The results indicate that the proposed scheme can perform better than Farid-Lyu's scheme and can attain a greater probability of correct decision at the same probability of false alarm. Another advantage of the proposed scheme is the reduced amount of training samples. While the SVM classification in the Farid-Lyu's scheme [11] needs sample tampered images under all types of manipulations for training, it is not necessary for the proposed technique as the decision is made just by comparing the estimated coefficients from the test image with a reference pattern (untampered image). Thus, the suggested method is more universal as it can more efficiently classify even tampering distortions that was not previously considered.

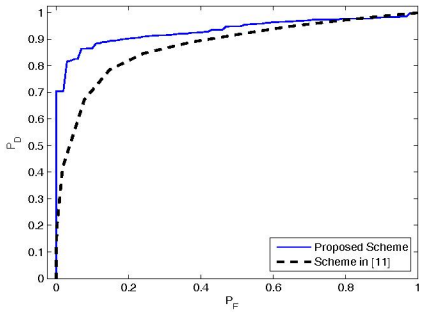

Fig. 6. Receiver operating characteristics

5. SUMMARY AND CONCLUSIONS

In this work, we introduce a new formulation to study the problem of image tampering. The proposed method considers the images captured by the camera are authentic and any further manipulations done to it would make it inauthentic. Based on an elaborate *ground-truth* modelling, we characterize the properties of an authentic camera output. We model the linear part of post-camera processing as a tampering filter and find its coefficients using blind deconvolution. The estimated filter coefficients are then used to identify post-camera processing, such as filtering, compression, rotation, etc. We show through detailed simulations that the proposed technique is efficient and does not require any prior knowledge about the nature of tampering operations.

6. REFERENCES

[1] J. Fridrich, "Image Watermarking for Tamper Detection," *Proc. of Intl. Conference on Image Processing (ICIP)*, vol. 2, pp. 404–408, Oct 1998.

[2] A. C. Popescu and H. Farid, "Exposing Digital Forgeries by Detecting Traces of Re-sampling," *IEEE Transactions on Signal Processing*, vol. 53, no. 2, pp. 758–767, Feb 2005.

[3] J. Lukas and J. Fridrich, "Estimation of Primary Quantization Matrix in Double Compressed JPEG Images," *Proc. of DFRWS*, Aug 2003.

[4] A. C. Popescu and H. Farid, "Statistical Tools for Digital Forensics," *6th Intl. Work. on Info. Hiding & LNCS*, vol. 3200, pp. 128–147, May 2004.

[5] A. C. Popescu and H. Farid, "Exposing Digital Forgeries in Color Filter Array Interpolated Images," *IEEE Transactions on Signal Processing*, vol. 53, no. 10, pp. 3948–3959, Oct 2005.

[6] I. Avcibas, S. Bayram, N. Memon, M. Ramkumar, and B. Sankur, "A Classifier Design for Detecting Image Manipulations," *Proc. of Intl. Conference on Image Processing (ICIP)*, vol. 4, pp. 24–27, Oct 2004.

[7] A. Swaminathan, M. Wu, and K. J. Ray Liu, "Component Forensics of Digital Cameras: A Non-Intrusive Approach," *Proc. of Conference on Information Sciences and Systems*, Princeton, NJ, Mar 2006.

[8] J. Adams, "Interaction between Color Plane Interpolation and other Image Processing Functions in Electronic Photography," *SPIE Cameras and Sys. for Electronic Photography & Scientific Imaging*, Feb 1995.

[9] D. Kundur and D. Hatzinakos, "A Novel Blind Deconvolution Scheme for Image Restoration using Recursive Filtering," *IEEE Trans. on Signal Processing*, vol. 46, no. 2, pp. 375-390, Feb 1998.

[10] G. R. Ayers and J. C. Dainty, "Iterative Blind Deconvolution Method and its Applications," *Optics Letters*, vol. 13, no. 7, pp. 547–549, 1988.

[11] H. Farid and S. Lyu, "Higher-Order Wavelet Statistics and their Application to Digital Forensics," *IEEE Workshop on Statistical Analysis in Computer Vision*, Madison, WI, 2003.