RFS for detecting Image Manipulation

[1]Farooq Sunar Mahammad, [2]N. Varshitha Reddy, [3]S.V. Navya Sahithi, [4] L.V. Kowsalya [5] V.Chinmayi
[1] Associate Professor, Department of CSE, Santhiram Engineering College, Nandyal, A.P
[2,3,4,5] IV.B.Tech, Department of CSE, Santhiram Engineering College, Nandyal, A.P.

ABSTRACT: The data driven image manipulation detection in the presence of attacker with limited knowledge about the detector. At first, we imagine that the attacker knows the architecture, training data and class of features that detector depend on. The analyst designs the detector that depends on subset of features chosen at random in the class in order to get an advantage in his race of arms with the attacker. Given its ignorance about the exact feature set, the attacker attacks a version of detector is based on the entire feature set. By this way the effectiveness of the attacker decreases, because there is no guarantee for the attacker that his attack is successful as the detector is working on full feature space. Thus, random feature selection increases the security of the detector for negligible loss of performance in the absence of attacks. We focus on two specific kinds of image manipulations, namely Adaptive histogram equalization and Median filtering.

Key Words: Detectors, Secret Key, Image Forensics, Feature Extraction.

I. INTRODUCTION: DEVELOPING secure image forensic tools, capable of granting good performance even in the presence of an adversary aiming at impeding the forensic analysis, turns out to be a difficult task, given the weakness of the traces the forensic analysis relies on[1]. As a matter of fact, a number of Counter-Forensics (CF) tools have been developed, whose application hinders a correct image forensic analysis. Early CF techniques were rather simple, as they consisted in the application of some basic processing operators like noise dithering, recompression, resampling or filtering. Though often successful, the application of general post processing operators, sometimes referred to as laundering, does not guarantee that the forensic traces are completely erased and hence does not necessarily result in the failure of the forensic analysis. When the attacker has enough information about the forensic algorithm, much more effective CF techniques can be devised. By following the taxonomy introduced in, we say that we are in a Perfect Knowledge (PK) scenario, when the attacker has complete information about the forensic algorithm used by the analyst. In the PK case, very powerful CF techniques can be developed allowing the attacker to prevent a correct analysis by introducing a limited distortion into the attacked image. Generally speaking, the attacker needs only to solve an optimization problem looking for the image which is in some sense closest to the image under attack and for which the output of the forensic analysis is the wrong one. Even if such an optimization problem may not be always easy to solve, the exact knowledge of the decision function allows the application of powerful techniques either in closed form, or by relying on gradient-descent mechanisms [2].

II. LITERATURE SURVEY:

2.1. Anti-forensics of contrast enhancement in digital images:

The authors Stamm, Matthew, and KJ Ray Liu proposed a blind forensics of contrast enhancement in digital images has attracted much attention of the forensic analyzers[3]. "Blind forensics of contrast enhancement in digital images." They discussed about new variants of contrast enhancement operators which are undetectable by the existing contrast enhancement detectors based on the peak-gap artifacts of the pixel gray level histogram. Local random dithering is introduced into the design of contrast enhancement mapping for removing such artifacts. Effectiveness of the proposed anti-forensic scheme is validated by experimental results on a large image database for various parameter settings. The developed anti-forensic techniques could verify the reliability of existing contrast enhancement forensic tools against sophisticated attackers and serve as the targets for developing more reliable & secure forensic techniques.

A wide variety of multimedia editing software's, both commercial and open source, are currently available to every computer user. The facility and powerful

editing functionality of such software's makes digital image manipulation become easy and frequent. So the originality, integrity and even authenticity of digital images may suffer destruction. To recover the human's trust on digital image data, there is an increasing need for developing techniques to detect digital image manipulation in the manner of blind and passive. Image manipulation forensics is just such a technique.

In general, prior works on digital image manipulation forensics can be labeled into two categories. In the first category, forensics methods concentrate on identifying the content-changing image manipulations including image splicing and copy-move, which reshape the image content visually and semantically. In the second category, content-preserving image manipulations such as resampling, compression, contrast enhancement, blurring, sharpening and median filtering are detected or estimated passively. Besides the wide application in the general image processing pipeline, the content-preserving manipulations are often used to conceal visual tampering trail and destroy the forensically significant statistical fingerprints. As a result, blind detection of the content-preserving operations is still significant. Recently, the blind detection and estimation of image contrast enhancement have been concerned extensively. In, the blind forensic algorithms for detecting the globally and locally applied contrast enhancement have been proposed. They perform contrast enhancement detection by seeking out unique peak-gap artifacts introduced into an image's histogram. In the recent paper, the authors propose an iterative algorithm to jointly estimate the contrast enhancement mapping used to modify an image and the pixel value histogram of the unenhanced image.

In contrast with cryptography, multimedia forensics remains an inexact new science without strict security proofs [4]. Although the existing forensic tools are good at uncovering naive manipulations in the scenario without attacks, there is a lack of awareness on their behavior in the practical application. Because little is known about the reliability of forensic techniques against a sophisticated counterfeiter, who is aware of the techniques in detail. Anti-forensic is just the technique employed by an image forger to hide or remove the forensically significant manipulation fingerprints, with the aim to deceive the forensic detectors. To the best of our knowledge, currently there are only three anti-forensic techniques: undetectable image re-sampling, synthesis of color filter array pattern and anti-forensic of JPEG compression.

The authors[5] propose a counter forensic method in the form of targeted attacks against the state-of-the-art contrast enhancement forensic algorithms. Local random dithering is introduced into the design of pixel value mapping for removing the histogram peak-gap artifacts. Simultaneously, the same visual enhancement effect as that of traditional contrast enhancement is still be preserved. Such an integrated tamper hiding method [6] can be easily tuned to the style of post-processing attack by adding random noise onto the traditional contrast-enhanced image.

2.2. Can We Trust Digital Image Forensics?

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

Back in analog times, a photograph was generally perceived as a "piece of truth". With digital image processing replacing its analog counterpart, critics have expressed the concern that it has never been so easy to manipulate images. The advent of low-cost digital imaging devices as well as powerful and sophisticated editing software makes it no longer necessary to obtain specialist skills to alter an image's tenor. Thus, questions regarding image

authenticity are of growing relevance, especially in contexts where nowadays multimedia society bases important decisions on them. Lately discovered forgeries in newspapers and scientific journal are only the tip of the iceberg. Particular attention has to be drawn to courtroom applications, in which the authenticity of photographs as pieces of evidence deserves utmost importance.

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

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

To draw attention to this disproportion, authors focuson two specific forensic methods — a resampling detector proposed by Popescu and Farid [7] and an approach to digital camera identification by Luk´aˇs, Fridrich and Goljan [8] and develops ways to deceive these methods. Forensic methods might benefit from research on countermeasures in a similar way as reasoning about attacks in multimedia security in general is useful to improve security. In this sense, attacks on image forensic algorithms can be understood as schemes to systematically mislead the detection methods.

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

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

2.3. Evasion Attacks Against Machine Learning at Test Time:

In security-sensitive applications, the success of machine learning depends on a thorough vetting of their resistance to adversarial data. In one pertinent, well-motivated attack scenario, an adversary may attempt to evade a deployed system at test time by carefully manipulating attack samples. In this work, we present a simple but effective gradient based approach that can be exploited to systematically assess the security of several, widely-used classification algorithms against evasion attacks. Following a recently proposed framework for security evaluation, we simulate attack scenarios that exhibit different risk levels for the classier by increasing the attacker's knowledge of the system and her ability to manipulate attack samples. This gives the classier designer a better picture of the classier performance under evasion attacks, and allows him to perform a more informed model selection (or parameter setting). We evaluate our approach on the relevant security task of malware detection in PDF _les, and show that such systems can be easily evaded. We also sketch some countermeasures suggested by our analysis.

Machine learning is being increasingly used in security-sensitive applications such as spam filtering, malware detection, and network intrusion detection. Due to their intrinsic adversarial nature, these applications differ from the classical machine learning setting in which the underlying data distribution is assumed to be stationary. To the contrary, in security-sensitive applications, samples (and, thus, their distribution) can be actively manipulated by an intelligent, adaptive adversary to confound learning; e.g., to avoid detection, spam emails are often modified by obfuscating common spam words or inserting words associated with legitimate emails. This has led to an arms race between the designers of learning systems and their adversaries, which is evidenced by the increasing

complexity of modern attacks and countermeasures. For these reasons, classical performance evaluation techniques are not suitable to reliably assess the security of learning algorithms, i.e., the performance degradation caused by carefully crafted attacks.

To better understand the security properties of machine learning systems in adversarial settings, paradigms from security engineering and cryptography have been adapted to the machine learning field. Following common security protocols, the learning system designer should use proactive protection mechanisms that anticipate and prevent the adversarial impact. This requires (i) finding potential vulnerabilities of learning before they are exploited by the adversary; (ii) investigating the impact of the corresponding attacks (i.e., evaluating classifier security); and (iii) devising appropriate countermeasures if an attack is found to significantly degrade the classifier's performance.

Two approaches have previously addressed security issues in learning. The min-max approach assumes the learner and attacker's loss functions are antagonistic, which yields relatively simple optimization problems. A more general game-theoretic approach applies for non-antagonistic losses; e.g., a spam filter wants to accurately identify legitimate email while a spammer seeks to boost his spam's appeal. Under certain conditions, such problems can be solved using a Nash equilibrium approach. Both approaches provide a secure counterpart to their respective learning problems; i.e., an optimal anticipatory classifier.

III.PROPOSED WORK:

Authors[12] propose to randomize the selection of the feature space wherein the analysis is carried out. To be specific, let us assume that to achieve his goal - hereafter deciding between two hypotheses H0 and H1 about the processing history of the inspected image - the analyst may rely on a large set Vof, possibly dependent, features. The number of features used for the analysis may be in the order of several hundreds or even thousands; for instance, they may correspond to the SPAM features described in [9] or the rich feature set introduced in [10]. In most cases, the use of all the features in Vis not necessary and good results can be achieved even by using a small subset of V. Our proposal to secure the forensic analysis is to randomize it by choosing a random subset of V– call it Vr- and let the analysis rely on Vronly; in a certain sense, the randomization of the feature space can be regarded as a secret key used by the analyst to improve the security of the analysis. Given its ignorance about the exact feature set used by the analyst, a possibility for the attacker is to attack the entire feature set V. Random feature selection increases the security of the detector for negligible loss of performance in the absence of attacks. Authors focus on two specific kinds of image manipulations, namely Adaptive histogram equalization and Median filtering.

3.1 Secure Detection By Random Feature Selection:

In this, we first describe the security assumptions behind our work, then we give a rigorous definition of binary detection based on Random Feature Selection (RFS)[11] and provide a theoretical analysis to evaluate the security vs robustness trade-off under a simple statistical model. Though derived under simplified assumptions, the theoretical analysis is an insightful one since it provides useful insights on the impact that the statistics of the host features has on the security of the randomized detector. In addition, it permits to analyze the dependence of classification accuracy on the number of selected features both in the presence and in the absence of attacks. Even if the paper focuses on RFS, the theoretical framework is a general one and can also be used to analyze other kinds of (linear) feature randomization.

3.2 Image Manipulation Detection:

The theoretical analysis given in the previous section suggests that a detector based on a randomized subset of features provides a better security with respect to a full-feature detector. The applicability of such an idea to real world applications, however, requires great care, since the assumptions behind the theoretical analysis are ideal ones and are rarely met in practice.

3.3 Attack in the Feature Domain:

In this section, we describe the feature domain attack we have implemented against the RFS SVM detectors.

IV.RESULTS AND ANALYSIS:

4.1. Adaptive histogram equilization:

Fig a: Original Image

The above figure represents the original image. When the attacker attacks the fig a then the image gets manipulated as shown in the fig b.

Fig b: Attacked Image

By the above fig b, it can be identified that the image is being attacked by Adaptive Histogram Equalization.

Fig c: Detected Image

After the image was attacked, the original image is detected by using the technique Random Feature Selection. A secret key is applied to detect the original image.

4.2. Median Filtering:

Fig e: Original Image

The above figure represents the original image. When the attacker attacks the fig a then the image gets manipulated as shown in the fig f.

Fig f: Attacked Image

By the above fig e, it can be identified that the image is being attacked by Median Filtering.

Fig g: Detected Image

After the image was attacked, the original image is detected by using the technique Random Feature Selection. A secret key is applied to detect the

original image.

V.CONCLUSION:

This is to exploit randomization, specifically feature space randomization, to restore the trust towards the quality of the forensic analysis in adversarial settings. From this theoretical implementation, we can use more accurate models to further reduce the gap between the analysis and the conditions encountered in real applications.

VI.FUTURE ENHANCEMENT:

The extension of our approach is to counter attacks against detectors based on deep learning, specifically convolution neural networks [13]. In such a case, in fact, the features used by the detector are not chosen by the analyst, since they are determined by the network during the training phase, hence calling for the adoption of other forms of randomization for some preliminary works in this direction.

VIII.REFERENCES:

[1]. T. Gloe, M. Kirchner, A. Winkler, and R. Bohme, "Can we trust digital image forensics?" in Proc. 15th ACM Int. Conf. Multimedia, Augsburg, Germany, Sep. 2007, pp. 78–86.

[2].Klein, S., Pluim, J.P., Staring, M. and Viergever, M.A., 2009. Adaptive stochastic gradient descent optimisation for image registration. International journal of computer vision, 81(3), p.227.

[3]. Stamm, Matthew, and KJ Ray Liu. "Blind forensics of contrast enhancement in digitalimages." 2008 15th IEEE International Conference on Image Processing. IEEE, 2008.

[4]. Böhme, R., Freiling, F.C., Gloe, T. and Kirchner, M., 2009, August. Multimedia forensics is not computer forensics. In International Workshop on Computational Forensics (pp. 90-103). Springer, Berlin, Heidelberg.

[5].Böhme, Rainer, and Matthias Kirchner. "Counter-forensics: Attacking image forensics." Digital image forensics. Springer, New York, NY, 2013. 327-366.

[6]. Kirchner, M. and Böhme, R., 2007, June. Tamper hiding: Defeating image forensics. In International Workshop on Information Hiding (pp.326-341). Springer, Berlin, Heidelberg.

[7]. Kirchner, M., 2008, March. On the detectability of local resampling in digital images. In Security, Forensics, Steganography, and Watermarking of Multimedia Contents X (Vol. 6819, p. 68190F). International Society for Optics and Photonics.

[8]. Lukas, Jan, Jessica Fridrich, and Miroslav Goljan. "Digital camera identification from sensor pattern noise." IEEE Transactions on Information Forensics and Security 1.2 (2006): 205-214.

[9]. T. Pevný, P. Bas, and J. Fridrich, "Steganalysis by subtractive pixel adjacency matrix," IEEE Trans. Inf. Forensics Security, vol. 5, no. 2, pp. 215–224, Jun. 2010.

[10]. J. Fridrich and J. Kodovský, "Rich models for steganalysis of digital images," IEEE Trans. Inf. Forensics Security, vol. 7, no. 3, pp. 868–882, Jun. 2012.

[11]. Chen, Z., Tondi, B., Li, X., Ni, R., Zhao, Y. and Barni, M., 2019. Secure detection of image manipulation by means of random feature selection. IEEE Transactions on Information Forensics and Security, 14(9), pp.2454-2469.

[12]. Gutub, Adnan, Ayed Al-Qahtani, and AbdulazizTabakh. "Triple-A: Secure RGB image steganography based on randomization." 2009 IEEE/ACS International Conference on Computer Systems and Applications. IEEE, 2009.

[13]. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W. and Abbeel, P., 2017, September. Domain randomization for transferring deep neural networks from simulation to the real world. In 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS) (pp. 23-30). IEEE.