FORENSICS VS. ANTI-FORENSICS: A DECISION AND GAME THEORETIC FRAMEWORK

Matthew C. Stamm, W. Sabrina Lin, and K. J. Ray Liu

Dept. of Electrical and Computer Engineering, University of Maryland, College Park

ABSTRACT

In order to combat the spread of digital forgeries, researchers have developed a variety of forensic techniques to verify the authenticity of digital multimedia files. Though many of these techniques can reliably detect traditional forgeries, recent research has shown that they can easily be fooled by anti-forensic operations designed to hide evidence of forgery. In response, new forensic techniques have been developed to detect the use of anti-forensics. In light of this, there is now a need to develop a theoretical understanding of the interactions between a forger using anti-forensics and a forensic investigator. In this paper, we propose techniques to evaluate the performance of anti-forensic algorithms along with a game theoretic framework for analyzing the interplay between forensics and anti-forensics. Furthermore, we propose a new automatic video frame deletion detection technique along with a technique to detect the use of video anti-forensics. We evaluate these techniques using our proposed analytical framework.

Index Terms— Anti-Forensics, Digital Forensics, Video Compression, Game Theory

I. INTRODUCTION

The prevalence of digital devices has caused digital multimedia content to become pervasive throughout modern society. However, because digital content can be easily altered using widely available software, its authenticity must be established before it can be trusted. As a result, a number of digital forensic techniques have been developed over the past decade to identify the origin of a media file, trace its processing history, and identify digital multimedia forgeries [1]. These techniques operate by identifying traces, known as fingerprints, left in digital multimedia files by manipulating operations or during the digital capture process.

Though digital forensic techniques are capable of detecting standard manipulations, recent research has shown that they can be fooled by a forger using *anti-forensics* to hide their forgery. Anti-forensic techniques operate by disguising manipulation fingerprints or falsifying device specific fingerprints inadvertently introduced when a digital file is formed. In previous work, we proposed an anti-forensic technique to remove compression fingerprints from digital images and showed how this technique can be used disguise several types of image forgery [2]. Additionally, we proposed an anti-forensic technique to hide evidence of frame deletion in digital videos [3]. Other anti-forensic operations have been designed to falsify the photo-response non-uniformity (PRNU) fingerprint left in digital images by sensor imperfections [4] and hide fingerprints left by image resizing or rotation [5].

The use of anti-forensics by a digital forger is not without its drawbacks, however. Many anti-forensic operations leave their own forensically identifiable fingerprints in digital multimedia files just like traditional signal processing operations. Researchers have already developed techniques to detect PRNU forgery [6] and identify the use of single JPEG compression anti-forensics [7].

This work is supported in part by AFOSR grant FA95500910179.

The authors can be reached by email at {mcstamm,kjrliu}@umd.edu.

In the past, the performance of digital forensic techniques has been measured using traditional tools from decision theory. While these tools can adequately evaluate forensic techniques, they often are poorly suited to measure the performance of anti-forensic operations. For example, should a missed forgery detection in an anti-forensically modified file be counted the same as one in which the file was not anti-forensically modified? If an anti-forensic operation is able to successfully remove fingerprints left by a particular forgery operation but introduces new fingerprints of its own, how do we evaluate its effectiveness?

In this last scenario, a forger may choose to reduce the strength of fingerprints left by their anti-forensic operation by decreasing the strength at which they apply anti-forensics. They must be careful, however, because this will cause a corresponding increase in the strength of the manipulation fingerprints that remain after anti-forensics has been used. The forensic investigator, meanwhile, must ensure that the combination of the false alarm rates from their techniques to detect editing and the use of anti-forensics is below a constant false alarm rate. As a result, the forger and forensic investigator must both balance a set of trade-offs that depend upon the actions of the other party. When examining these trade-offs, one may ask what are the optimal set of actions for both the forger and forensic investigator to take?

In this paper, we address these problems by proposing a set of techniques to evaluate the performance of anti-forensic operations. Additionally, we propose a game theoretic framework to evaluate the dynamics between a forger and a forensic investigator. This framework can be used to determine the probability that a forgery will be detected when both a forger and forensic investigator are using optimal anti-forensic and forensic detection strategies. We then demonstrate the usefulness of these techniques by evaluating a set of video forensic and anti-forensic techniques with them. To do this, we propose an automatic frame deletion detection technique. This technique improves upon Wang and Farid's method which requires human inspection [8]. Additionally, we propose a new forensic technique to detect the use of our anti-forensic frame deletion method. Using our proposed framework, we are able to determine under which conditions a video forgery will likely be detected.

II. PERFORMANCE ANALYSIS OF ANTI-FORENSICS

Consider the forensic problem of determining if a digital multimedia file ψ has been manipulated using an editing operation $m(\cdot)$. Traditionally, this is posed as a hypothesis testing problem where the null hypothesis is that ψ is unaltered and the alternate hypothesis is that ψ is a manipulated version of another multimedia file ψ', i.e.

$$\begin{aligned} H_{0m} &: \psi \neq m(\psi'), \\ H_{1m} &: \psi = m(\psi'). \end{aligned} \tag{1}$$

We use the subscript m to differentiate this hypothesis test from other hypothesis testing problems which we will discuss later.

A forensic investigator will decide between these hypotheses using a decision rule δ_m. Typically, this decision rule operates by obtaining some measure of the strength of the fingerprints left in ψ by m, then comparing this measure to a decision threshold.

This decision threshold is chosen to maximize the decision rule's probability of detection $P_d(\delta_m) = P(\delta_m = H_1|\psi = m(\psi'))$ without exceeding a constraint on its probability of false alarm $P_d(\delta_m) = P(\delta_m = H_1|\psi = m(\psi'))$. We denote a decision rule designed with the false alarm constraint P_{fa} as $\delta_m^{(P_{fa})}$.

To disguise their forgery, a digital forger can design an anti-forensic operation α_m to fool the detector δ_m. In the past, the performance of α_m has been measured by the probability that δ_m will classify an anti-forensically manipulated file unmanipulated, or explicitly $P(\delta_m(\alpha_m(\psi)) = H_0|\psi = m(\psi))$. This measure can be misleading, however, because it unfairly attributes all missed detections to the anti-forensic operation. In reality, unless δ_m is able to perform with $P_d = 100\%$, it will naturally miss several manupulation detections even if anti-forensics is not used. Because of this, this measure is biased towards overestimating the performance of α_m.

To more accurately measure the performance of an anti-forensic operation, we propose using its *probability of anti-forensic effectiveness*, which we define as

$$P_{ae}(\alpha_m) = P(\delta_m(\alpha_m(\psi)) = H_0|\, \delta_m(\psi) = H_1, \psi = m(\psi)). \tag{2}$$

This measure avoids the previously discussed bias towards overestimating an anti-forensic operation's performance.

If, however, forensic investigators are aware of the existence of α_m, it does not need to achieve a $P_{ae} = 100\%$ in order to render δ_m ineffective. Instead, it only needs to reduce the performance of δ_m to the point that it provides investigators no advantage over making a completely random decision. This is equivalent to reducing δ_m's probability of detection to $P_d(\delta_m^{(P_{fa})}) = P_{fa}$. If this can be done, we claim that δ_m is susceptible to the anti-forensic attack α_m because decisions made by δ_m convey no information about whether ψ has been manipulated or not.

To measure the degree to which a forensic technique δ_m operating with a false alarm constraint P_{fa} is susceptible to an anti-forensic attack α_m, we define its *anti-forensic susceptibility* as

$$S_\alpha(\delta_m, P_{fa}) = \frac{P_d(\delta_m^{(P_{fa})}) - \max(P_d(\delta_m^{(P_{fa})})(1 - P_{ae}(\alpha_m)), P_{fa})}{P_d(\delta_m^{(P_{fa})}) - P_{fa}}. \tag{3}$$

The anti-forensic susceptibility is a measure between 0 and 1 of the decrease in effectiveness of δ_m caused by α_m. If $S_\alpha(\delta_m, P_{fa}) = 0$, this indicates that α_m is not able to cause any decrease in the performance of δ_m. Alternately, $S_\alpha(\delta_m, P_{fa}) = 1$ signifies that the performance of δ_m has been reduced to that of a random decision.

In order to provide an intuitive understanding of the anti-forensic susceptibility, we note that the numerator of S_α is the decrease in δ_m's probability of detection caused by the use of anti-forensics. This corresponds to the distance A in Fig. 1. When calculating this decrease, we take the maximum between probability that δ_m will detect manipulation if anti-forensics is used, i.e. $P_d(\delta_m^{(P_{fa})})(1 - P_{ae}(\alpha_m))$, and the P_{fa} because a forensic investigator can always achieve $P_d = P_{fa}$ by randomly deciding that a multimedia file is manipulated with probability P_{fa}.

The denominator of S_α is chosen to be the maximum decrease in δ_m's probability of detection that α_m needs to cause in order to render δ_m ineffective. This distance is denoted by B in Fig. 1. By choosing the denominator in this manner, we are able to normalize the measure S_α. We note that S_α is undefined at $P_{fa} = 100\%$ because no decrease in the performance of δ_m is possible at this false alarm level (δ_m will classify every file as manipulated).

III. TRADE-OFF BETWEEN FORENSICS AND ANTI-FORENSICS

When an anti-forensic operation leaves behind its own unique fingerprints, a new forensic detection technique δ_α can be designed

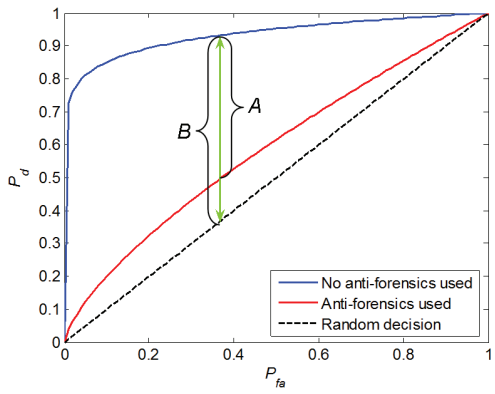

Fig. 1. Example relating the anti-forensic susceptibility to a forensic technique's ROC curves when anti-forensics is and is not used. The anti-forensic susceptibility at a given false alarm rate is the ratio A/B.

to detect the use of anti-forensics. As before, this detection problem can be framed as the following hypothesis testing problem

$$\begin{aligned} H_{0\alpha} &: \psi \neq \alpha_m(m(\psi')), \\ H_{1\alpha} &: \psi = \alpha_m(m(\psi')). \end{aligned} \tag{4}$$

Here δ_α is used to determine if the multimedia file in question ψ is a manipulated and anti-forensically modified version of another file ψ'.

If a forensic investigator is able to identify the use of anti-forensics in a file, they will logically assume it is inauthentic. This poses a difficult problem for a forger: if the use of anti-forensics can be detected, should it be used to disguise a forgery? Logically, a forger should use anti-forensics only if it decreases the probability that their forgery will be detected. In many cases, a forger can adjust the strength with which they apply anti-forensics. By applying anti-forensics with decreased strength, the forger can decrease the strength of the fingerprints left by anti-forensics. Care must be taken when doing this because as the strength of an anti-forensic operation is decreased, the strength of manipulation fingerprints that remain in the multimedia file after anti-forensics is applied will increase. In addition, the forger must take into account the cost of any perceptual distortion caused by anti-forensics because if the forgery does not appear perceptually realistic, it will be flagged as inauthentic. As a result, the forger must determine the optimal strength with which to apply anti-forensics.

A similar trade-off exists for a forensic investigator. Typically, a forensic investigator must operate with a constraint on their probability of false alarm. Since both the manipulation and anti-forensics detection techniques will contribute to the total probability of false alarm, the forensic investigator must choose a set of decision thresholds for δ_m and δ_α such that the total probability of false alarm lies within their constraint. When doing this, they can choose to allow one detection technique to operate with a higher P_{fa} than the other as long as the total false alarm constraint is met. This will increase the probability of detection achieved by the detection technique operating at a higher P_{fa} while lowering the probability of detection for the other. A rational forensic investigator will seek out the combination of thresholds that maximizes the total probability that they identify a forgery.

When examining these trade-offs, it is clear that the optimal anti-forensic strength used by the forger depends on the decision thresholds used by δ_m and δ_α. Similarly, the forensic investigator's optimal choice of decision thresholds for δ_m and δ_α depends on the strength with which the forger applies anti-forensics. The dependence of each party's actions upon those of the other naturally leads to the following question: does there exist a set of actions

that neither party has an incentive to deviate from? If such a set of actions exists, what are the probabilities of detection and false alarm achieved by the forensic investigator? To answer these questions, we propose the following game theoretic formulation for analyzing the dynamics between a forger and a forensic investigator.

Let player 1 denote the forensic investigator and player 2 denote the forger. We adopt the convention that the forensic investigator chooses the decision thresholds for δ_m and δ_α first, then allows the forger to choose their optimal anti-forensic strength in response.

Since choosing the set of decision thresholds for each forensic technique is equivalent to choosing the probabilities of false alarm at which they operate, a set of actions, also known as a strategy, for the forensic investigator can be completely specified by choosing the probability of false alarm η at which δ_m operates. For a given constraint ξ on the total probability of false alarm P_{fa}^{Tot}, the corresponding false alarm probability $\tilde{\eta}$ at which δ_α operates is given by finding $\tilde{\eta}$ such that $P_{fa}^{Tot} = \xi$. The total probability of false alarm is explicitly defined as

$$P_{fa}^{Tot} = P\left(\delta_m^{(\eta)}(\psi) = H_{1m} \bigcup \delta_\alpha^{(\tilde{\eta})}(\psi) = H_{1\alpha} \,\middle|\, \psi \neq m(\psi'), \psi \neq \alpha_m(m(\psi'))\right). \quad (5)$$

The set of possible strategies that the forensic investigator can employ is $\eta \in [0, \xi]$. Let $\alpha_m^{(k)}$ denote an anti-forensic operation operating at strength $k \in [0, 1]$ where $k = 1$ corresponds to using anti-forensics at full strength. The set of strategies that the forger can employ is the set of anti-forensic strengths $k \in [0, 1]$.

Given a pair of strategies (η, k), we define the utility that player 1 wishes to maximize as

$$U_1(\eta, k) = P\Big(\delta_m^{(\eta)}(\psi) = H_{1m} \bigcup \delta_\alpha^{(\tilde{\eta})}(\alpha_m^{(k)}(\psi)) = H_{1\alpha} \Big| \psi = m(\psi')\Big). \quad (6)$$

This is the probability that a forgery will be detected, either by detecting evidence of manipulation or the use of anti-forensics. By contrast, player 2 wishes to minimize this quantity along with some measure $\gamma(\cdot)$ of the perceptual distortion introduced into their forgery by the use of anti-forensics. As a result, the utility of player 2 is

$$U_2(\eta, k) = -U_1(k, \eta) - \gamma\Big(m(\psi), \alpha_m^{(k)}(m(\psi))\Big). \quad (7)$$

If closed form expressions for the probabilistic quantities that define U_1 and U_2 are known, then the Nash equilibrium strategies (η^*, k^*) can be analytically derived using standard techniques. If one player operates at their Nash equilibrium strategy, the other player gains no advantage by choosing any other strategy, thus both players have no incentive to deviate from the Nash equilibrium strategies. If no closed for expression for these utilities exist, the Nash equilibria can be determined numerically.

After the Nash equilibrium strategies have been identified, the probability that the forensic investigator detects a forgery is given by evaluating $U_1(\eta^*, k^*)$. Since both the Nash equilibrium strategies and probability of forgery detection are influenced by the forensic investigator's false alarm constraint ξ, these quantities will likely change as ξ is varied. By determining the probability of forgery detection at the Nash equilibrium for each $\xi \in [0, 1]$, a new ROC curve can be constructed showing the forensic investigator's ability to detect forgeries if both players act rationally. We define this ROC curve as the *Nash equilibrium receiver operating characteristic curve*, or NE ROC curve.

IV. EXAMPLE USING VIDEO FORENSICS AND ANTI-FORENSICS

To provide an example of how our proposed evaluation techniques apply to real forensic scenarios, we demonstrate them on the problem of video forensics and anti-forensics.

In many scenarios, a video forger may wish to delete a sequence of frames from a digital video. This may be done to hide evidence of a particular event. In prior work, Wang and Farid demonstrated that a forensically detectable fingerprint is left in MPEG videos by frame deletion [8]. This fingerprint takes the form of periodic spikes in the total motion prediction error in the video's P-frames. Wang and Farid proposed detecting this fingerprint by visually inspecting the sequence of total motion prediction errors in a video's P-frames for periodic spikes.

Recently we proposed an anti-forensic technique capable of preventing frame deletion fingerprints from occurring in forged videos [3]. It operates by increasing the total motion prediction error for each P-frame to the level of these spikes in prediction error so that they are no longer detectable. If the total motion prediction error of a P-frame does not correspond to a spiky value, we set several of that P-frame's motion vectors to zero then recalculate its motion prediction error. Since setting the motion vectors to zero will result in a poorly predicted frame, the motion prediction error will increase.

We note that since both the anti-forensically modified motion vectors and the associated prediction error are used to reconstruct the frame during MPEG decoding, this anti-forensic technique will introduce essentially no distortion into the video. This is because even after anti-forensic modification, a frame is still equal to the sum of the motion prediction version of the frame and its prediction error. A detailed explanation of both frame deletion fingerprints and our anti-forensic technique can be found in [3].

Here we propose an automatic technique for detecting frame deletion. It operates by first median filtering the sequence of total motion prediction errors in each P-frame to obtained a smoothed version. Next this smoothed prediction error sequence is subtracted from the actual prediction error sequence, and the DFT of the resulting signal is calculated. If frames have been deleted from or added to the video, a peak will occur in the frequency bin $k = N/T$ where N is the length of the sequence of P-frame prediction errors and T is the number of P-frames that the video encoder places in each group of pictures. To perform detection, we measure the strength of this peak and compare it to a decision threshold.

Additionally, we propose a technique to detect the use of video anti-forensics. This technique exploits the fact that even though anti-forensic modification sets several of a video's motion vectors to zero, the true motion between video frames is unchanged. It operates by first decompressing a video, then estimating its motion vectors as if we wished to re-encode it. For each P-frame, we calculate the mean Euclidean distance between our estimated motion vectors and the motion vectors contained in the compressed video. If the video has been anti-forensically modified, the Euclidean distance between the motion vectors will be large for frames that have been anti-forensically modified. Otherwise, the estimated motion vectors will closely match those used during compression. As a result, we create a feature vector containing the mean Euclidean distance between both sets of motion vectors across all frames, along with a measurement of the periodicity of the sequence of distance between both sets of motion vectors. We use principal component analysis to reduce the dimensionality of this vector to a single dimensional feature and perform detection by comparing this feature to a decision threshold.

V. EXPERIMENTAL RESULTS

To evaluate the performance of each of these forensic and anti-forensic techniques, we compiled a database of 21 standard uncompressed video sequences such as the 'Foreman' and 'Carphone' sequences. Next we simulated MPEG-2 compression and decompression in Matlab and used this simulation to compress each video in our database. We then decompressed each video, deleted a number of frames from it, and recompressed each video both with and without using anti-forensics. We tested each of the resulting videos for evidence of frame deletion and the use of anti-forensics.

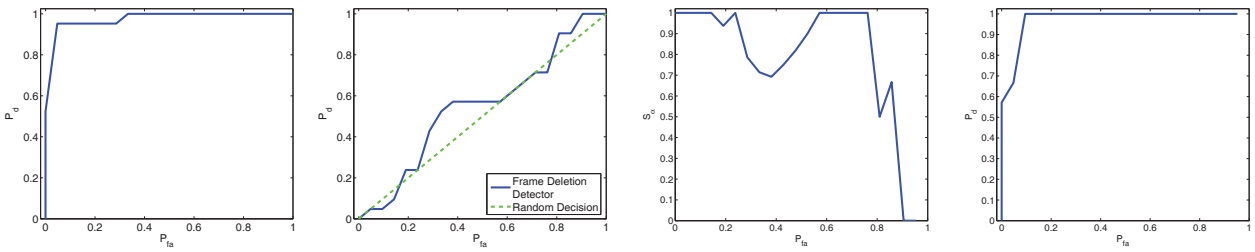

Fig. 2. Experimental results showing (a) ROC curve for our proposed frame deletion detection technique, (b) ROC curve for frame deletion detection if anti-forensics is used, (c) plot showning the anti-forensic susceptibility of our frame deletion detector to our anti-forensic technique, and (d) ROC curve for our proposed anti-forensics detection technique.

Fig. 2(a) shows an ROC curve displaying the performance of our automatic frame deletion detection technique when tested on videos that have not been anti-forensically modified. As can be seen in this figure, frame deletion can be detected with aP_d of 95% at a false alarm rate less than 5%. This indicates that if anti-forensics is not used, frame deletion can be detected very accurately.

Fig. 2(b) shows the ROC curve for our automatic frame deletion detection technique if anti-forensics is used. We can see from this figure that the performance of the detector is severely degraded if anti-forensics is used. This is emphasized by the results displayed in Fig. 2(c), which show the anti-forensic susceptibility of our frame deletion detector to our anti-forensic attack. These results show that for all $P_{fa} \leq 80\%$, our anti-forensic technique acheived an anti-forensic susceptibility of .7 or greater. Furthermore, for all $P_{fa} \leq 20\%$, the frame deletion detector performs no better than a random decision if anti-forensics is used.

An ROC curve displaying the performance of our video anti-forensics detection technique is shown in Fig. 2(b). These results show that if anti-forensics is used at full strength, a $P_d = 100\%$ can be achieved for a $P_{fa} \geq 10\%$.

After we measured the nominal performance of each of our forensic and anti-forensic techniques, we used our game theoretic framework to determine the probability of forgery detection at Nash equilibrium. To do this, we modified our anti-forensic technique to operate at variable strengths by making the anti-forensic increase in each P-frame's prediction error adjustable. We then modified each video with several different anti-forensic strengths and performed frame deletion and anti-forensics detection as before.

Because no distortion is introduced into the video by our anti-forensic technique, the term $\gamma(\cdot)$ in (7) can be set to zero. As a result, $U_2(k, \eta) = -U_1(k, \eta)$, thus reducing the trade-off between video forensics and anti-forensics to a zero sum game. This allowed us to find the Nash equilibrium strategies for a range of constraints on the forensic investigator's P_{fa}^{Tot} by using our experimental results to numerically solve the equation $(k^*, \eta^*) = \arg\max_\eta \min_k U_1(k, \eta)$. We used this data to create the NE ROC curve displayed in Fig. 3.

From this curve we can see that if the forensic investigator must operate with a P_{fa}^{Tot} constraint of 5% or less, frame deletion forgeries are difficult to detect. We note that for $P_{fa}^{Tot} \leq 5\%$, the Nash equilibrium probability of forgery detection is less than the P_d achieved by both the frame deletion detector and the anti-forensics detector at the same false alarm level. This reinforces the notion that the forger can create a more successful anti-forensic attack by decreasing its strength. If the forensic examiner is able to relax their P_{fa}^{Tot} constriant to 10% or greater, they will be able to detect frame deletion forgeries at with at least 85% probability.

VI. CONCLUSIONS

In this paper, we have proposed a set of methods to evaluate the effectiveness of anti-forensic techniques. Additionally, we have proposed a game theoretic framework to analyze the dynamic interplay between a forger and a forensic investigator. We have introduced the Nash equilibrium ROC curve as a method of evaluating the ability of a forensic investigator to detect a forgery when both they and a forger are operating using optimal forensic and anti-forensic strategies. We have proposed new forensics to detect frame deletion in digital videos and to detect the use of frame deletion anti-forensics. We have analyzed video frame deletion forensics and anti-forensics using our proposed game theoretic framework and shown that frame deletion forgeries are difficult to detect for $P_{fa}^{Tot} \leq 5\%$ but can be reliably detected for $P_{fa}^{Tot} \geq 10\%$.

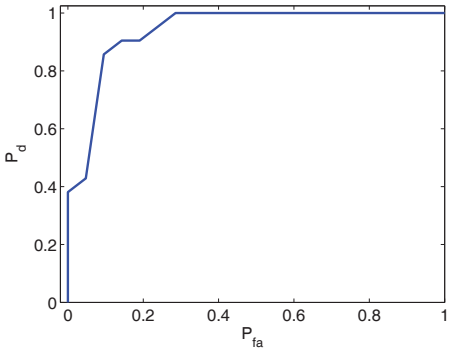

Fig. 3. Nash equilibrium ROC curve for video frame deletion detection.

VII. REFERENCES

[1] H. Farid, "Image forgery detection," *IEEE Signal Processing Magazine,*, vol. 26, no. 2, pp. 16 –25, Mar. 2009.

[2] M.C. Stamm and K.J.R. Liu, "Anti-forensics of digital image compression," *IEEE Trans. Information Forensics and Security*, vol. 6, no. 3, pp. 1050 –1065, Sep. 2011.

[3] M. C. Stamm and K. J. R. Liu, "Anti-forensics for frame deletion/addition in MPEG video," in *Proc. IEEE ICASSP*, Prague, Czech Republic, May 2011, pp. 1876 – 1879.

[4] T. Gloe, M. Kirchner, A. Winkler, and R. Böhme, "Can we trust digital image forensics?," in *15th Int. Conf. Multimedia*, 2007, pp. 78–86.

[5] M. Kirchner and R. Bohme, "Hiding traces of resampling in digital images," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 4, pp. 582–592, Dec. 2008.

[6] M. Goljan, J. Fridrich, and Mo Chen, "Defending against fingerprint-copy attack in sensor-based camera identification," *IEEE Trans. Inf. Forensics and Security*, vol. 6, no. 1, pp. 227 –236, march 2011.

[7] G. Valenzise, V. Nobile, M. Taglisacchi, and S. Tubaro, "Countering JPEG anti-forensics," in *Proc. IEEE ICIP*, Brussels, Belgium, Sep. 2011.

[8] W. Wang and H. Farid, "Exposing digital forgeries in video by detecting double MPEG compression," in *ACM Multimedia and Security Workshop*, Geneva, Switzerland, 2006.