Peer-reviewed version available at *Sensors* **2019**, *19*, 2424; doi:10.3390/s19112424

Article

Detecting Malicious False Frame Injection Attacks on Video Surveillance at the Edge using Electrical Network Frequency Signals

Deeraj Nagothu [1], Yu Chen [1,*], Erik Blasch [2], Alexander Aved [2] and Sencun Zhu [3]

[1] Dept. of Electrical and Computer Engineering, Binghamton University, Binghamton, NY 13902, USA; {dnagoth1, ychen}@binghamton.edu
[2] The U.S. Air Force Research Laboratory, Rome, NY 13441, USA; {erik.blasch.1, alexander.aved}@us.af.mil
[3] Dept. of Computer Science and Engineering, Penn State University, University Park, PA 16802, USA; szhu@cse.psu.edu
* Corresponding author: ychen@binghamton.edu

Abstract: Over the past few years, the importance of video surveillance in securing the national critical infrastructure has significantly increased, whose applications include detecting failures and anomalies. Accompanied by video proliferation is the increasing number of attacks against surveillance systems. Among the attacks, false frame injection (FFI) attacks that replay video frames from a previous recording to mask the live feed has the highest impact. While many attempts have been made to detect FFI frames using features from the video feeds, video analysis is computationally too intensive to be deployed on-site for realtime false frame detection. In this paper, we investigate the feasibility of FFI attacks on compromised surveillance systems at the edge and propose an effective technique to detect the injected false video and audio frames by monitoring the surveillance feed using the embedded Electrical Network Frequency (ENF) signals. An ENF operates at a nominal frequency of 60Hz/50Hz based on its geographical location and maintains a stable value across the entire power grid interconnection with minor fluctuations. For surveillance system video/audio recordings connected to the power grid, the ENF signals are embedded. The time-varying nature of the ENF component is used as a forensic application for authenticating the surveillance feed. The paper highlights the ENF signal collection from a power grid creating a reference database and ENF extraction from the recordings using conventional short-time Fourier Transform and spectrum detection for robust ENF signal analysis in the presence of noise and interference caused in different harmonics. The experimental results demonstrate the effectiveness of ENF signal detection and/or abnormalities for FFI attacks.

Keywords: Video Surveillance; Visual Layer Attack; Electrical Network Frequency (ENF) Signal; False Frame Injection (FFI) Attack.

1. Introduction

Physical infrastructure security and human safety rely on surveillance systems to monitor activities with minimal human intervention. A common example is audio-video systems for detecting human trespassing [1]. Some methods also provide safety by alerting first responders with emergent events to improve safety [2], [3]. On the other hand, the proliferation of smart surveillance systems has made them attractive to physical-layer, network-based visual data attacks [4]. These attacks are primarily designed to compromise the audio-video feed to disguise malicious activities or prevent detection. Among them, visual data attacks are a special dimension that only exists in video surveillance systems [5].

Frame duplication attacks is one type of visual data attacks. It pre-records idle events and upon triggering, replays the pre-recorded video and audio frames to mask the current events. Frame duplication attacks result in compromised alarms, which are solely dependent on the surveillance feed

Peer-reviewed version available at Sensors 2019, 19, 2424; doi:10.3390/s19112424

received. Even with human intervention to monitor the surveillance data, the malicious activity can go unnoticed. An attacker's actions could be catastrophic in case of government or banking infrastructure break-ins, where physical security has extremely high priority. Many algorithms have been proposed to detect frame duplication or mirroring attacks [6], but most of these detection techniques are performed on previously stored media files that can be delayed from event occurrence.

With the proliferation of edge computing and the Internet of Things (IoT) technology, Smart Cities envision the public safety surveillance as an edge service. The capability of instant, on-site detection of visual layer attacks, i.e., false frame injection attacks (FFI), becomes essential to keep cities and communities safe [7].

The *Electric Network Frequency* (ENF) is an instantaneous frequency in power distribution networks, which varies across its nominal frequency 50/60Hz based on the power supply demand from the consumers. The fluctuation in ENF is typically desired to be close to the nominal frequency [8], and the deviation of ENF from its nominal frequency in the United States is between 59.90Hz and 60.10Hz, whereas in Asian and European countries it is between 49.90Hz and 50.10Hz. The instantaneous behavior of the ENF is useful because the fluctuations are the same within a power grid. The instantaneous values of varying power supply frequency across the nominal frequency are represented as the ENF signal. It has been observed that the surveillance feed contains traces of ENF in both audio and video recordings. The source of ENF in video recording is a light source, like a fluorescent lamp, and in case of audio recording, it could be either from the electromagnetic field interference, mechanical vibrations of electrical powered devices, or the audible hum from powered devices [8].

In this paper, we propose an online authentication system using ENF signal to quickly detect the malicious false frame injection attacks (also referred to as *frame duplication attacks* or *replay attacks*). Specifically, our work is focused on the ENF signal extracted from the audio recordings from the surveillance feed due to its high reliability and efficiency as compared to video recordings that need a powered light source [9]. The embedded ENF traces are extracted using signal processing techniques like STFT (Short Time Fourier transform), which exploit the presence of ENF signals in multiple harmonics [10]. To establish the extracted signal reliability, the ENF signal is collected directly from the power supply and stored as a reference database. The database includes ENF signal variation w.r.t time and zone of extraction. The major contributions of this work are:

- The feasibility of frame duplication attacks at the edge has been investigated, and an attack with smart adaptability to environment and automatic triggering mechanism is implemented and tested;
- The authenticity of the ENF signals is validated using signal traces collected at multiple locations within the same power grid;
- A robust method is proposed to extract the fluctuations in the audio recordings and to compare with the reference ENF power signal using the cross-correlation factor;
- The relationships between the strength of the acoustic mains hum and the signal to noise ratio (SNR) of the ENF signal are verified; and
- The effectiveness and correctness of the proposed detection scheme are validated through an experimental study using real-world ENF signal traces.

The rest of the paper is organized as follows. Section 2 provides the background knowledge of ENF and the related work regarding the attacks on a surveillance system. Section 3 illustrates the feasibility of launching a frame duplication attack at the edge through an actual implementation. Section 4 introduces our method to detect false frame injection attacks utilizing the ENF signals embedded in the recorded audio and provides available techniques on video recordings. Section 5 presents the experimental results that verify the effectiveness of the proposed method. Section 6 concludes this paper along with a brief discussion regarding our future work.

Peer-reviewed version available at Sensors 2019, 19, 2424; doi:10.3390/s19112424

2. Background Knowledge and Related Work

2.1. Attacks on a Surveillance System

Nowadays, video surveillance systems are arguably the most popular measure for the safety and security of physical facilities and residents of our communities. The emergence of more sophisticated attack tools and methods has brought deep concerns to researchers and stakeholders. Network-based attacks like Cross-site Scripting, buffer overflow, SQL Injection, and boot loader or firmware attacks give privileged access to unauthorized people. Gaining root access allows attackers to impair the normal function of a surveillance system by conducting more attacks, such as blinding cameras, disabling video sensors, eavesdropping, as well as data exfiltration and infiltration oriented visual-data layer attacks. These suspicious activities could escape from detection, and the attacker may even gain the command and control over the surveillance network [5].

In this paper, we focus on *data infiltration based visual-data layer attacks*. Frame duplication attacks are one of the most frequently encountered forgery attacks on a live video feed. Once the attacker has gained access over the surveillance cameras through network attacks, the attack code can control the surveillance output. By inserting previously recorded video and audio frames with normal scenarios, the on-going suspicious activities, personnel, or objects may go undetected. Many methods have been proposed to detect the replay attacks using the spatial and temporal domain similarities by extracting features from the video frames and analyzing these features to detect the frame forgery [11]. These algorithms mostly extract features from a video sub-sequence and compare them with other sub-sequences for similarity [12]. A number of correlation techniques [6], [13], [14] have also been adopted to identify frame duplication and region duplication in a video. All these similarity detection techniques require a stored surveillance recording database, and hence they require much computation time to process each video frame. In the case of surveillance systems, the late discovery of such forgery after the event does not afford intervention, incident capture, or property anti-theft. Real-time detection and alarm indication is a top priority.

In order to launch a false frame duplication attack, the attack code works in a controlled environment. It is recognized that the environmental factors change continuously, like the light intensity of the surrounding due to daytime or nighttime, an object's position in the point of view (POV) of the camera, or the introduction of new objects [7]. If there are visible differences between the pre-recorded frames used for attack and the current genuine frames, the security personnel may beware of it immediately. Hence the attack code continually looks for any change in the camera's POV and updates the pre-recorded frames with the new changes made in the environment. The environment monitoring allows the attacker to always have up-to-date recorded frames which can be triggered at any instant. For example, using a simple facial recognition software in the attack code, an attacker can launch the attack upon detecting a specific face (or as simple as a quick response (QR) code). In this paper, for demonstration we will use a face detection based trigger to launch the attack, and collect the surveillance feed for analysis (discussed further in section 3).

2.2. Electrical Network Frequency Signals

ENF signals can be extracted using various techniques from both audio and video recordings. The collection of ENF signals is also affected by many factors including the environment of recording, and the recording device itself. Initially, ENF traces were found in recorders that were directly connected to a power grid, and other researchers showed that ENF signals are also present in battery powered devices [15], [16]. The source of ENF in such battery powered devices is the audible hum from any electrical device running on power from the main grid and generating noise, where the noise carries the time-varying nature of the ENF traces [16], [17]. For battery-powered devices, a device in motion can have high noise and interference caused due to air friction in the ENF frequency zone and hence making ENF extraction more difficult [18].

Peer-reviewed version available at Sensors 2019, 19, 2424; doi:10.3390/s19112424

In this paper, ENF signals are extracted from audio recording made by the surveillance cameras, which are connected to the power grid. Audio signals are recorded at a sampling rate of 8 KHz. This sampling rate provides room to capture the ENF traces in multiple harmonics including the nominal frequency of 50/60Hz and consumes less storage. Meanwhile, the high video frame rate of surveillance cameras makes it difficult to capture the ENF that varies with high time resolution. Some earlier research has extracted the ENF signal by capturing changes in light intensity using optical sensors, aliasing frequency, rolling shutter, and a super-Pixel based approach [9], [19], [20]. However, these techniques are computation intensive, which makes them not practical for edge devices.

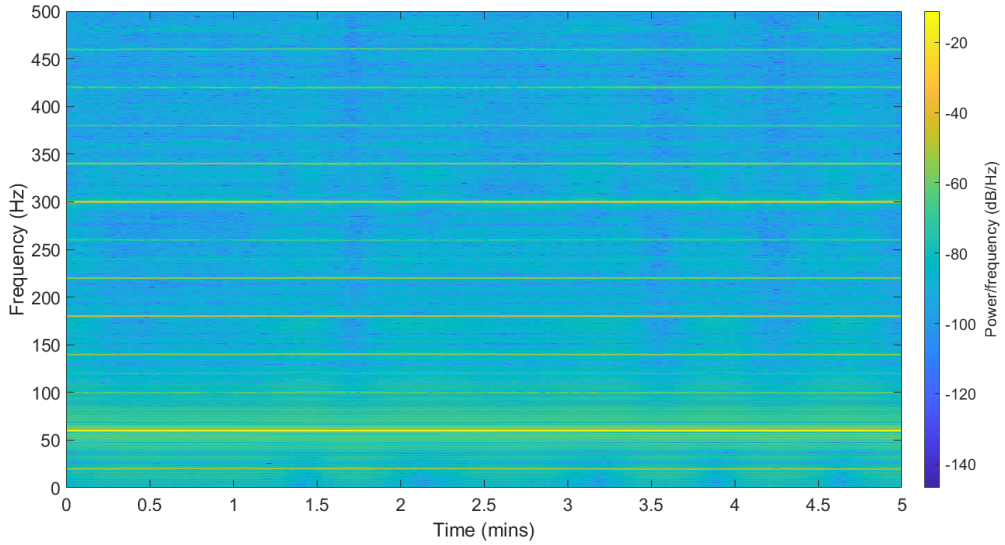


Figure 1. Spectrogram of Power recording.

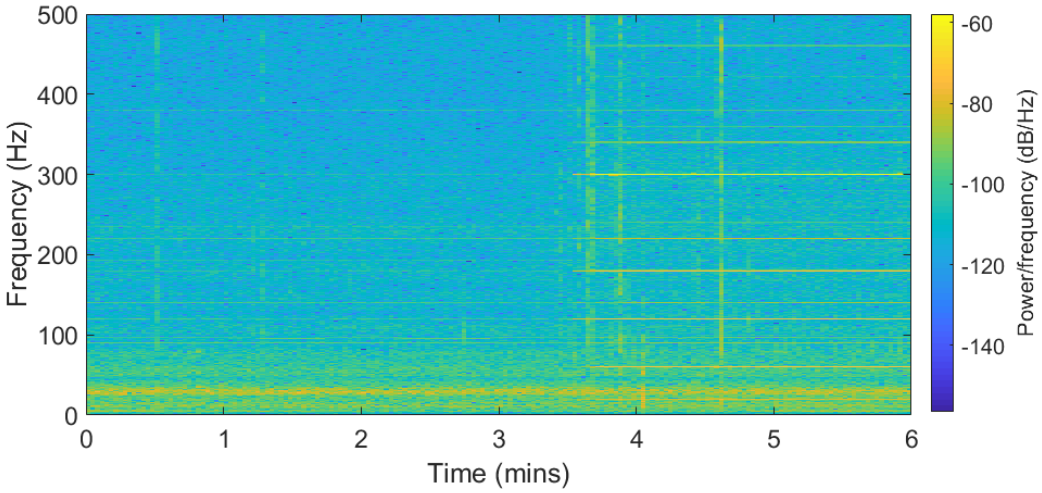


Figure 2. Spectrogram of Audio recording with noise source after 3.5 minutes.

The ENF signals can be collected using a circuit consisting of a step-down transformer and a voltage divider circuit. Figure 1 shows a spectrogram of the collected power recordings with the ENF traces embedded in it. The Signal to Noise (SNR) ratio is high around the nominal frequency zone. The example is recorded at Binghamton University in the United States, so the nominal ENF frequency is around 60Hz and it varies in the range of $\pm$0.02Hz. The range of variation changes per the location; for instance, India and Lebanon have a frequency variation around the nominal frequency in the range

Peer-reviewed version available at Sensors 2019, 19, 2424; doi:10.3390/s19112424

±0.8Hz. The ENF variations are observed to appear in many harmonic bands along with the nominal frequency band [10]. These harmonic bins have different signal strengths as compared to the nominal bin.

Figure 2 represents the audio recording spectrogram. The recording was made in an android phone connected to the power supply for six minutes, where the first 3.5 minutes were recorded without main power electrical devices like computers or speakers operating nearby. Then, after 3.5 minutes the electrical devices around the recorder were powered on. The ENF traces are available in the nominal frequency along with its harmonics for the second part of the recording. In the first part of the recording, the ENF traces are captured as a result of direct power grid connection. It is also possible that the traces were captured due to low-energy ambient noise from devices running farther away from the recorder. The recordings show that ENF traces can be captured in the presence of acoustic hum or from devices directly connected to the power grid.

2.3. ENF Signal Applications

ENF signals have been adopted in digital forensics to authenticate digital media recordings [8], [17]. The use of ENF technique was firstly demonstrated to authenticate media recordings as proof for legal jurisdiction purposes to verify whether or not an evidence was tampered with. The ENF authentication technique was introduced and multiple extraction processes have been discussed [8]. Many forgeries as false evidences were detected using the instantaneous ENF signal. Robust extraction of the ENF signals has been an active research topic, and multiple signal extraction and tracking algorithms have been proposed [8], [21], [22], [23]. The signal extraction experiments on alternating current (AC) powered recording devices and battery powered devices reveal the source of ENF in a battery powered device is the acoustic hum generated by the electrical devices connected to a main power source [16], [17]. These experiments show that the main power noise source in the proximity of the recording devices can result in capturing ENF traces.

A high precision phase analysis technique was introduced, which checks for sudden changes in the phase and amplitude of the extracted ENF signal [23]. This technique does not rely on a pre-built reference database, but there were cases where the deleted or added video clip could have the same phase as the proceeding clip. Hence, there are not any observable phase or amplitude changes to utilize. As the ENF signals are embedded in multiple harmonics along with the nominal frequency range, a multi-estimator model could enable a more robust extraction of ENF signals from a weak spectral component [10], [21]. The estimator model states that the frequency variations of the harmonic spectral range have larger variance compared to the nominal frequency. It has also been observed that for different types of recording environments, recording devices with different microphones like dynamic or electric microphones result in ENF traces with high SNR in specific harmonic ranges as compared to rest of the spectra. The extraction process includes combining multiple spectral frequency ranges, resulting in a robust signal with a low computational requirement.

These previous studies demonstrate the usefulness of the ENF, so we adopt this technique to extract the ENF signals from our surveillance recordings. Various environmental factors and device-related scenarios, like wave interference, Doppler effect and movement of the recording device with respect to the noise source, could affect ENF capture in audio [18]. For instance, due to the different types of microphones used, the ENF signals are embedded in multiple harmonics. Figures 3 and 4 represent two ENF instances recorded at the same time in two different rooms and different buildings. The ENF signals are very similar throughout a power grid, and the slight shift might be due to the oscillator error in two different device recorders.

Algorithms to extract the ENF signals from video recordings along with the audio samples can be developed simultaneously. For example, ENF traces can be detected in video recordings using optical sensor measurements with indoor lighting [9]. A light source was required during the video recordings, and the availability of ENF traces in surveillance camera video recordings made using CCD camera sensors were confirmed using frequency aliasing. An alternative approach to extract ENF fluctuations

Peer-reviewed version available at Sensors 2019, 19, 2424; doi:10.3390/s19112424

from CMOS camera recordings uses rows from each video frame leveraging a rolling shutter technique [24]. This technique cannot be universally applied to all cameras since the number of rows exposed to light in each image sensor changes with the manufacturers. A Super-Pixel based approach divides a video frame into a group of pixels with similar pixel intensity as known as *super-pixels* [20]. The instantaneous light condition variations in these super-pixels are used to detect the presence of ENF in a given video file without investing a lot of processing power and time on video files with no ENF traces.

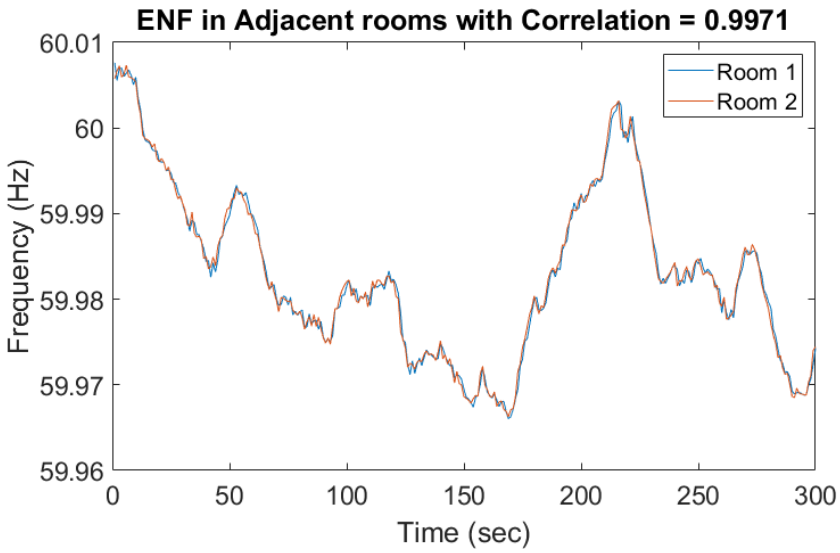


Figure 3. ENF captured in Adjacent rooms

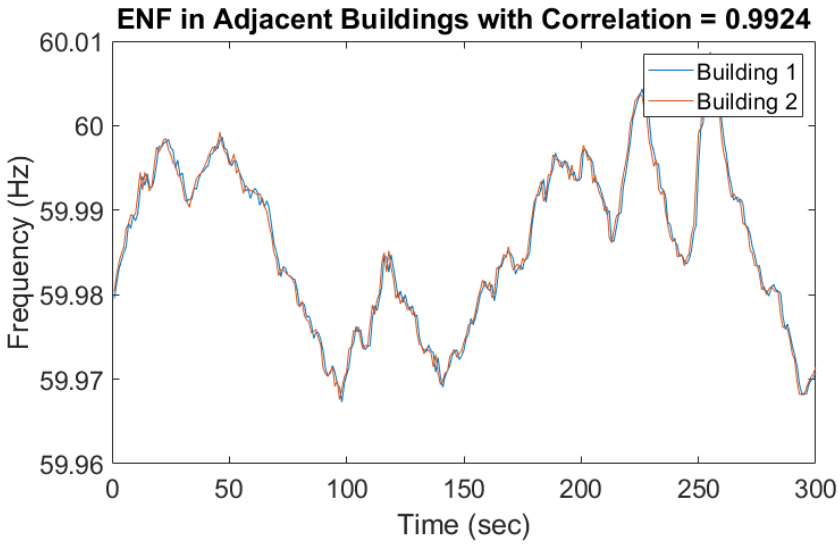


Figure 4. ENF captured in Adjacent Buildings

3. Real-time Frame Duplication Attack Implementation

Before introducing our ENF based detection mechanism, this Section investigates the feasibility of an automated real-time frame duplication attack at the network edge by an experimental case study. The constructed attacking system also serves as the test-bed for the detection scheme validation.

Peer-reviewed version available at Sensors 2019, 19, 2424; doi:10.3390/s19112424

3.1. Overview

To launch a real-time frame duplication attack, we assume that the edge based surveillance systems have been compromised through network attacks. This allows the attacker to gain complete access to the live video feed along with the manipulation of the output stream as required. The algorithm devised includes two modules, monitoring for audio-video replay and deploying an attack.

Figure 5 represents the algorithm flow diagram. In the first module "monitoring audio-video replay", it consists of collecting duplicate recording in two parallel processes where video and audio streams are monitored independently. The video monitoring process constantly checks for any motion in the frame, and when a static scene is detected, an automated recording of the static scene is started in the background process. The motion detection algorithm in the video process performs Gaussian blur on the frames to smooth out edges and minimize errors due to noise, and then changes in pixel intensities are compared with a threshold to detect any motion. The audio monitoring process detects noise in the environment and records audio samples when there is no background noise. With the "monitoring audio-video replay", a recent recording of the video and audio are collected and stored. The second module "deploying an attack" represents detecting a trigger and launching the attack. The mechanism used as a cue is the face recognition algorithm. When the trigger event is detected, the video frames and audio samples are combined and deployed to mask the live video feed.

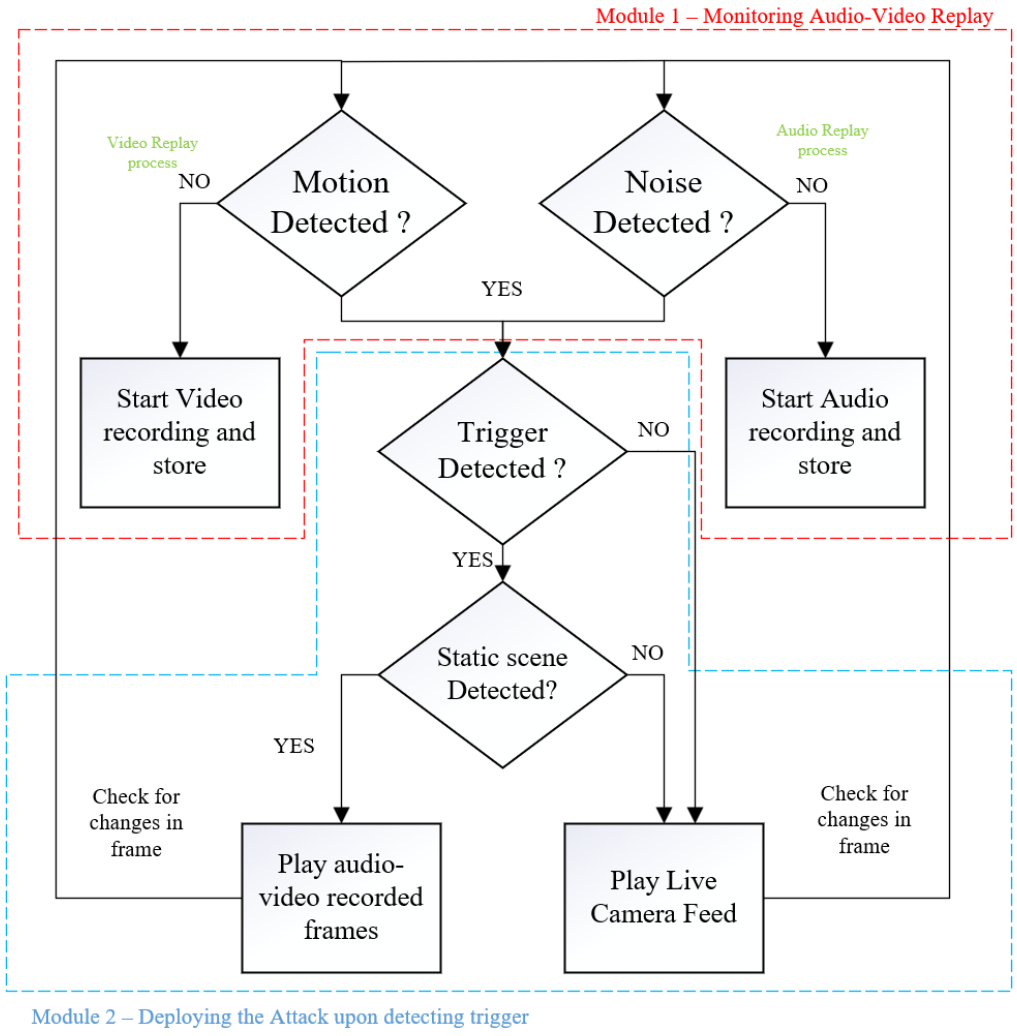


Figure 5. Flow diagram of the frame duplication algorithm

Peer-reviewed version available at Sensors 2019, 19, 2424; doi:10.3390/s19112424

3.2. Attack Algorithm Functionality

The monitoring audio-video replay module discussed earlier consists of two parallel processes, for video and audio, running independently to collect replay recording. The term *replay recording* represents a pre-recorded video frames or audio samples to be used later by the algorithm when the attack is triggered. The motion detection algorithm in the video process is used to detect an occurrence of a static scene by comparing the pixels in consecutive frames. The changes in pixel intensity are compared with a threshold, where different environments have different sensitivity to pixel changes and hence different threshold values.

For our test-bed, we consider indoor environments where the changes in pixel values are more stable compared to outdoor environments. The changes which occur indoors are people walking, gradual changes in natural light intensities and artificial light changes. The algorithm is tuned to detect these changes in the frames by using a Gaussian blur on incoming frames. The Gaussian blur performs convolution on the image, acting as a low pass filter and therefore attenuating high-frequency components more than the lower-frequency components. Since human movement in the camera view appears as a low-frequency change while noise is a high-frequency change, removing the noise helps the algorithm better distinguish human motions from noise. Below is the Gaussian function for calculating the transformation to apply to each pixel in the image

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

where σ^2 is the variance of the Gaussian distribution, and x and y are the distances from the origin in the horizontal axis and the vertical axis, respectively.

In visual replay attacks, a duplicated streaming video out of synchronization with its audio could potentially raise suspicions to people monitoring the surveillance. Hence, the second parallel process, where the audio process is running to detect static noise in the environment and collect audio replay recording. For example, when a static video is replayed in the live feed if the audio in the background has surrounding noise which is independent of video, then it would raise suspicions. So, the video frames and audio samples are recorded independently and replayed together to represent a static scene with no background noise. The audio replay recording is collected when there is no noise detected. A Fast Fourier Transform (FFT) is performed on the samples to obtain a frequency domain representation of the input audio stream. Noise detection is performed by taking the mean volume across all frequencies and comparing it to a threshold. The threshold for audio is also decided based on different environmental settings of the camera.

The trigger detection in the "deploying attack module" is responsible for detecting a pre-determined event and using the audio-video replay recording as pseudo-live feed. In this paper, face detection (of the attacker) module is used as a triggering event. For modern surveillance cameras, a high-quality video stream is captured with decent frames per second (FPS) compared to the surveillance cameras a decade ago. For the face detection module, the FPS processed is lower, but a single frame with the required face model detected is enough to trigger the attack and makes the processing speed irrelevant. In the face detection module, we use Histogram of oriented gradients (HOG) for fast human/face detection [25], [26]. The gradients of human faces are trained using a machine learning algorithm, where each face has a unique encoding. The perpetrator's face encoding was generated beforehand and embedded in the algorithm. When the perpetrator shows up in the camera view, the encoding vector is detected, and this event is used as a trigger mechanism for the replay attack. To avoid suspicions by deploying the attack as soon as the face is detected, the attack is instead placed on hold until a static scene appears again, and then the frames are replayed to mask the live feed.

The face detection model is used as an example to demonstrate the remote triggering capabilities of malicious algorithms. The trigger mechanism could also be performed manually using a command and control server to communicate with all the compromised surveillance cameras, or using a naturally occurring event to leave no traces of attacker appearing in the frames. Other examples of triggering

Peer-reviewed version available at Sensors 2019, 19, 2424; doi:10.3390/s19112424

event could be a specially designed QR-code on a T-shirt, unique hand gesture or even voice-activated trigger.

Figure 6 shows the frames observed by the camera (i.e., "live feed") and frames captured or delivered by the camera (i.e., "duplicated feed") when the attack is launched. In Fig. 6.a, the face encoding of a user (i.e., perpetrator) has been stored in the algorithm. When the perpetrator enters the scene, the camera detects the face along with other faces in the scenario. The perpetrator could walk into the scene with a group or individually, as long as the camera can detect the face and match it with the embedded face encoding. The HOG encoding is unique for different face structures, and hence it is faster to deploy facial recognition algorithm at the edge. Once a static scene is detected, the duplicated frames are replayed. Here, we opt to deploy the attack once a static scene appears again instead of immediately launching the attack. Deploying the attack with static scene avoids suspicious artifacts like sudden disappearance of a person from frame, and detecting duplicated frames in a static scene is harder than frames with objects in motion [14]. In Fig. 6.b, the periodic changes in the environment is reflected in the replay recording. The algorithm checks for changes in environment every two seconds and updates the stored recording accordingly. The second column represents the recording stored for future deployment of attack. The duration of the recording made is also modified based on the indoor-outdoor requirements. The capability of the attack algorithm to adapt to the changes in real-time shows the reliability of the algorithm in fooling a human perception and reduce suspicious behavior when a camera does not reflect the changes according to the environment. For example, a replay recording made at noon is used at night time; this can easily raise suspicion and alarm the authorities.

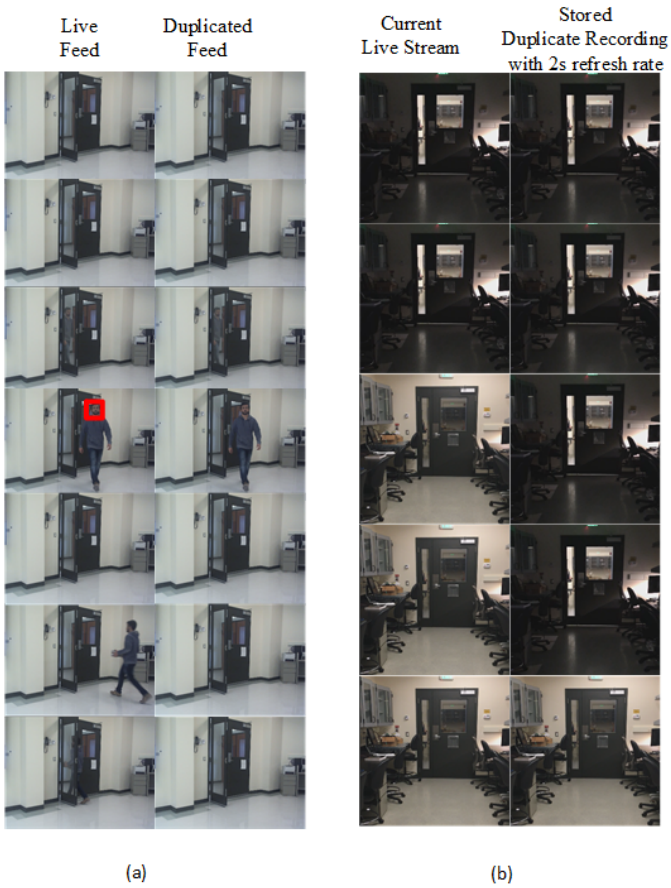


Figure 6. Frames of live and duplicated feed. (a) Attack triggered after detecting a face encoding and launched when a static scene appears (b) Updating the duplicated static scene recording based on changes in light intensity of the environment

Peer-reviewed version available at Sensors 2019, 19, 2424; doi:10.3390/s19112424

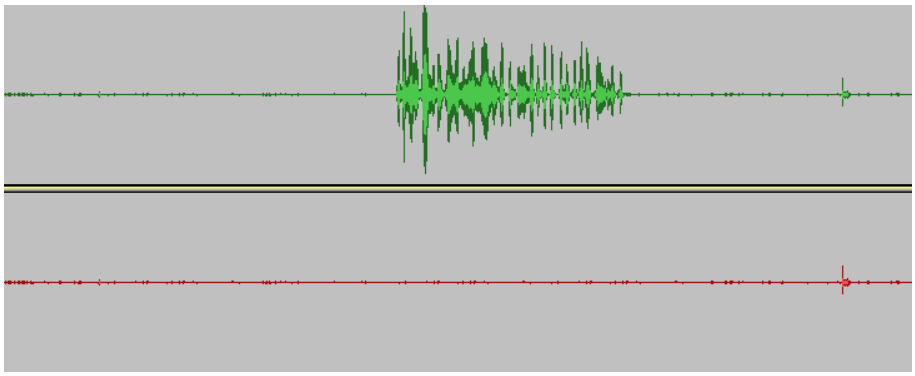


Figure 7. Original audio masked by replay recording of noiseless background recording. The top recording represents the original audio recording and the bottom recording is the duplicated recording after the attack is launched.

Along with the video frame duplication, the audio samples are also masked. Figure 7 represents masking noise made during the replay attack with its pre-recorded audio samples with no or less background noise. The allowed noise depends on the threshold used to compare the frequencies in FFT. For an indoor application, the noise level is assumed to be minimal, so higher frequency noise is eliminated from the replay recordings.

4. Detecting Malicious Frame Injection Attacks using ENF Signals

Inspired by the characteristics of the ENF signals, this work explores the feasibility of applying it to detect malicious frame injection attacks at the edge. In order to obtain a reliable ENF signal from the surveillance systems, we opt to use audio records as the source, which is insensitive to light conditions. A reliable database for authenticating the extracted ENF is created utilizing robust extraction techniques like the spectral combination of multiple harmonics. A *correlation coefficient threshold based* method is introduced to detect the existence of duplicated frames inserted by the attacker.

4.1. Applied Model

ENF traces occur around the nominal frequency range 50/60 Hz as $f_{ENF} = f_o + f_\Delta$, where f_o is the nominal frequency and f_Δ is the instantaneous frequency fluctuations from the nominal value. For power recordings, Fig. 1 shows the ENF traces at odd multiples of harmonic, with a strong signal at 60Hz. In case of audio recording, Fig. 2 shows that the traces occur more around even harmonics depending upon the type of microphone used.

For the spectrogram calculation of the recorded signal, we used a frame size of 1 second and nFFT = 8192, which gives a frequency resolution of 0.122 Hz for a signal with a sampling rate of 1000 Hz. The length of recorded signal used for each instance is six seconds. The power spectral density (PSD) of the ENF carrying signal is used to extract certain spectral bands $s(f)$, where the PSD $S(\omega)$ is computed from the fast Fourier transform (FFT) of the signal and $f \in k[f_o - f_v, f_o + f_v]$. f_v is the variation width of the ENF signal, f_o is the nominal frequency and k represents the harmonic frequency band.

The PSD $S_{N_{XX}}(f)$ is

$$S_{N_{XX}}(f) = \frac{1}{N}|X_N(f)|^2$$

where $X_N(f)$ is the Fourier transform of the signal

$$X_N(f) = \sum_{n=-\infty}^{\infty} x_n e^{-j\omega nT}$$

Peer-reviewed version available at Sensors 2019, 19, 2424; doi:10.3390/s19112424

where $w = 2\pi f$, T is the period of the signal duration, and n is the number of samples $1 \leq n \leq N$. Sampling at discrete times $x_n = x(n\Delta t)$ for a period $T = N\Delta t$, the PSD is

$$\overline{S}_{XX}(\omega) = \frac{(\Delta t)^2}{T} \left| \sum_{n=1}^{N} x_n e^{-j\omega n \Delta t} \right|^2$$

From the obtained spectral band, the instantaneous frequency for each frame window used is estimated by the maximum value in each power density vector obtained for that time instant. The period of signal duration represents the number of vectors obtained from PSD and instantaneous ENF values. Quadratic interpolation is used to obtain its dominant frequency from the maximum value in each vector. In Quadratic interpolation of the spectral peak, the peak location is given as

$$\Delta = \frac{1}{2} * \frac{\alpha - \gamma}{\alpha - 2 * \beta + \gamma}$$

where α is the previous bin of the max spectral bin, β is the max spectral peak and γ is the next bin. If k^* is the bin number of the largest spectral sample at the peak, where $1 \leq k^* \leq K$ for K bins, then $k^* + \Delta$ is the interpolated peak location of the bins and the final interpolated frequency estimate is

$$f_\Delta = (k^* + \Delta)\frac{f_s}{N}$$

here f_s is the sampling frequency and N is the number of FFT bins used. The instantaneous frequency estimate of the ENF signal is then given as $f_{ENF} = f_o + f_\Delta$.

4.2. Robust Extraction of ENF signals

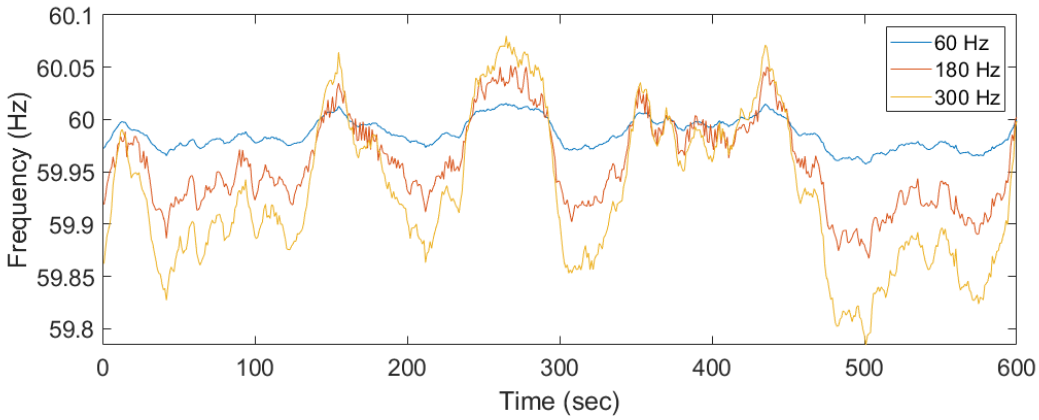


Figure 8. Different Harmonics of Power Recording shifted to 60Hz for comparison.

ENF traces appear in different harmonics with increasing frequency variations at different spectral bands. Figure 8 shows similar ENF fluctuations at odd/even harmonics. The power recordings are not affected by any noise since it is directly extracted from the power outlet, but in case of audio recordings, external noise could be captured and interfere with the ENF frequency ranges. The noise could lead to an inaccurate estimate of the ENF signal. A more robust technique was proposed to combine the spectral frequency bins from different harmonic bins based on the SNR [10]. The SNR is represented as the weight of spectral band, computed by the mean of the PSD in the ENF frequency range to the mean of spectral bin of that harmonic frequency.

$$w_k = \frac{\sum_{k=1}^{L} s(f_o - f_c, f_o + f_c)}{\sum_{k=1}^{L} s(f_o + f_c, f_o + f_v) + s(f_o - f_c, f_o - f_v)}$$

Peer-reviewed version available at Sensors 2019, 19, 2424; doi:10.3390/s19112424

where f_c is the range of ENF variations, and it is typically 0.02Hz in US and varies in European and Asian countries. f_v is the spectral band of interest in each of the k harmonics and f_o is nominal frequency. The weight obtained from each spectral bin is normalized and combined with different spectral bins to compute a combined spectrum of all harmonics containing ENF.

$$S(f) = \sum_{k=1}^{L} w_k s(f)$$

The normalized weight represents the SNR of harmonic frequency in different bands. The noise in some frequency band can be eliminated for the spectral bands with very low SNR. The approach is computationally more intensive for edge devices; therefore, a fog node is used to perform a second pass on ENF estimation on the audio recordings with more robust extraction by eliminating the false alarms produced by the edge devices. The discussion of the edge-fog-cloud hierarchy is beyond the scope of this paper – interested readers may find the architecture description in our related publications [27], [28], [29].

4.3. Correlation Coefficient for Extracted ENF Signals

ENF signal estimated from both power recording and audio recording for a small duration are compared to check for the similarity using a correlation coefficient between the two signals [30]. The ENF signal from power P_{ENF} and audio A_{ENF} is given as,

$$\rho(l) = \frac{\sum_{n=1}^{N}[f_{P_{ENF}}(n) - \mu_{P_{ENF}}][f_{A_{ENF}}(n-l) - \mu_{A_{ENF}}]}{var(P_{ENF}) * var(A_{ENF})}$$

$f_{P_{ENF}}$ and $f_{A_{ENF}}$ are the frequency estimation of the ENF signal from power and audio recordings, respectively. μ and var are the mean and variance of the frequency signal. l is the lag between the two signals. Even though the recordings are made at the same time, due to the oscillator error between the two devices the signals are not in sync. The lag is used to match the signals and a threshold decides the similarity between the two signals. If the difference between the reference and the current detection goes beyond a certain threshold, the system considers that a false frame injection attack is detected.

5. Experimental Results

5.1. Testbed Setup

A Raspberry Pi Model B is used as an edge device where the surveillance system is operating. An additional module with a sound card is added to record the power recording at the same time as the audio recordings. Python based code is used for the implementation and estimation algorithm of the ENF signal. The Python's parallel threading enables capturing and estimating the power ENF and the audio ENF simultaneously. The recordings are stored as a file in the common database. A laptop is used as a fog node to estimate the same ENF signals to verify the signal correlation in the second pass. Power recordings are made using a step-down transformer and a voltage divider circuit [8] and given as an input through 3.5mm audio jack. To reserve the computational power, the recordings are made in mono channel instead of a stereo channel. The signals are recorded at the sampling frequency of 8 KHz and it is down sampled to 1000 Hz for estimating the signals.

5.2. Implementation and Results

Both power and audio recordings with ENF traces are made simultaneously and the estimated ENF signals are compared based on the correlation coefficient obtained. We have implemented a visual-data layer replay attack and collected both the original and attacked audio recording along with the power recording simultaneously. Strong ENF traces were observed at 300Hz for both power and audio recordings. Figure 9 presents the estimated ENF from the power, original audio and attacked

Peer-reviewed version available at Sensors 2019, 19, 2424; doi:10.3390/s19112424

audio recording, respectively. The attacked audio includes pre-recording a selected period of time and it is replayed to mask the current original recording. The attack is launched at 300 seconds and a clear deviation between the original recording (green signal) and the attacked recording (red signal) can be observed. The part of the recording which is replayed is clearly seen from the signal comparison as the ENF estimates do not match, which indicates the possibility of forgery attacks on previously recorded media files. The correlation coefficient between the power ENF signal and the attacked ENF signal will be lower for the replayed part of the recording. Figure 9 conceptually validates the idea that ENF traces to distinguish an anomaly incurred by the injected frames.

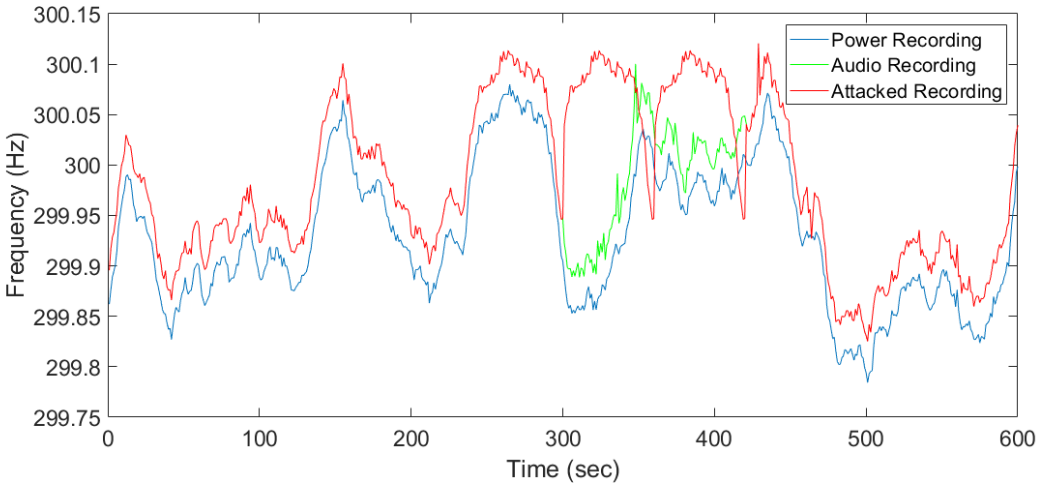


Figure 9. ENF estimated from the power and original audio recording.

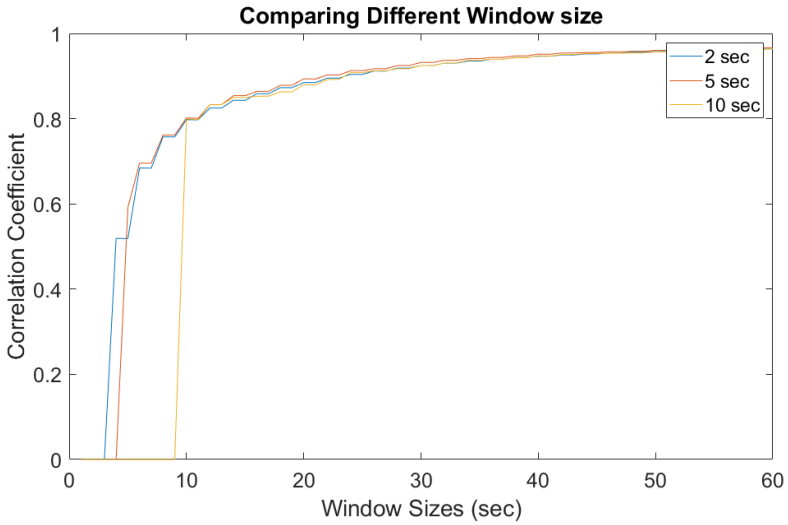


Figure 10. Window Sizes for Power and Original Audio.

In practice, a responsive surveillance system has to provide alerts instantly rather than help discern the problem from a delayed forensic analysis. Therefore, a sliding window based approach is introduced to extract and estimate the ENF from online records. A thorough study has been conducted for better understanding on different setup and overlap times between each ENF estimates. Comparisons were made with the correlation coefficient between those estimated ENF signals. Figures 10 and 11 show different window sizes used at the initial process. Based on the comparison between different shifting step lengths, it is clear that a window size of 25-30 seconds is the minimum to obtain a constant correlation coefficient of 0.8 and this value can be used as a threshold to detect dissimilar

Peer-reviewed version available at Sensors 2019, 19, 2424; doi:10.3390/s19112424

ENF signal estimations. Figure 10 is the correlation coefficient between the power signal and original audio signal. Figure 11 is estimated between power and attacked audio signal. It is clear that the correlation is higher for original audio signal compared to attacked audio.

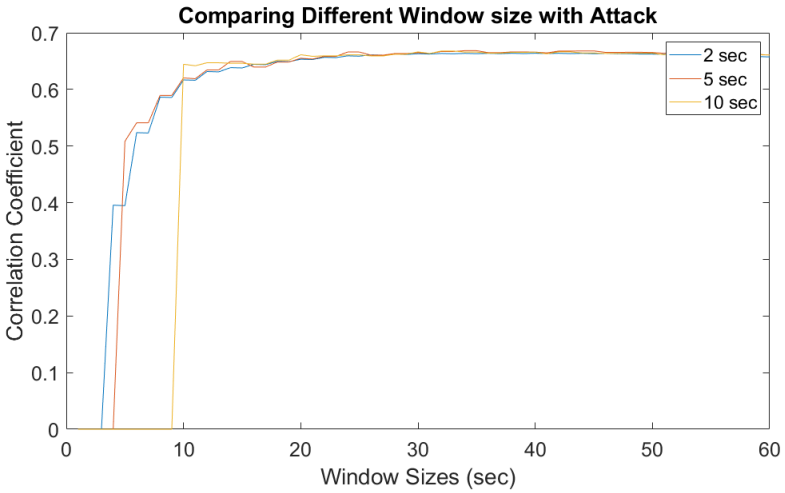
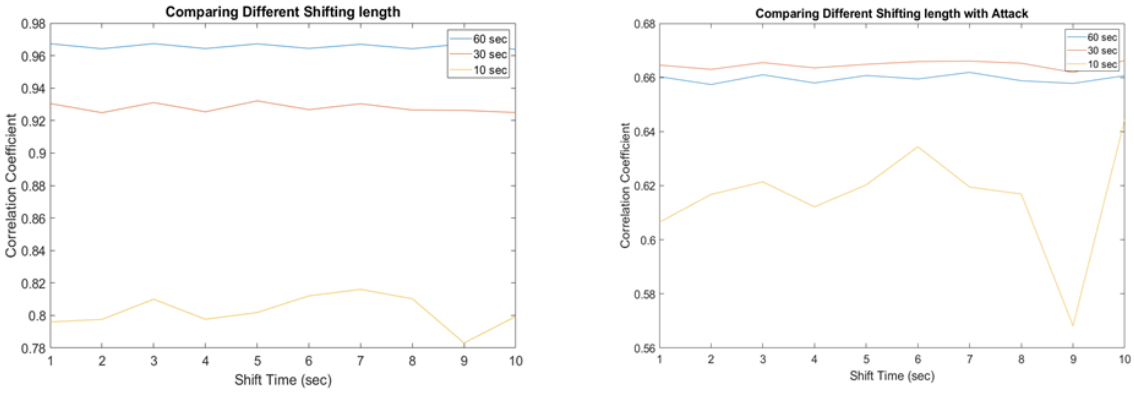


Figure 11. Window Sizes for Power and Attacked Audio.

Figures 12 is a detailed comparison between different window sliding step sizes. It is clear that with a smaller step size a higher correlation coefficient value is obtained compared to larger step sizes. However, the experimental study also shows that the computational overhead is higher with the smaller sliding window step sizes. A balanced point is that a window step size of five seconds allows a real-time response in case of mismatching signals. Taking multiple factors into consideration, our experimental results suggest a threshold for a correlation coefficient between two signals to be 0.8. A correlation coefficient above the threshold value of 0.8 means the video/audio stream is normal, while below 0.8 implies the possible existence of injected false frames. The lower the value is, the higher probability of attack.



(a) Shifting Lengths For Power and Original Audio

(b) Shifting Lengths For Power and Attacked Audio

Figure 12. Shifting Lengths comparing Power with Original and Attacked Audio.

Figure 13 is the comparison between different window sizes with a sliding window shift step size of five seconds. Even though the window size of ten seconds has smaller initialization delay, it is susceptible to a high false positive rate. The fluctuations in the correlation can be seen for original recording where some windows are not similar. In case of a 30 second or 60 second window size, the detection of frame duplication attack is similar. The 60 second window has less fluctuations between adjacent windows and the threshold of 0.8 clearly separates the distribution of duplication attacking

Peer-reviewed version available at Sensors 2019, 19, 2424; doi:10.3390/s19112424

scenarios with the actual normal recording. Comparing with Fig. 14, which represents a window shift step size of ten seconds, it is clear that the shift step size has lower impact compared to the window sizes.

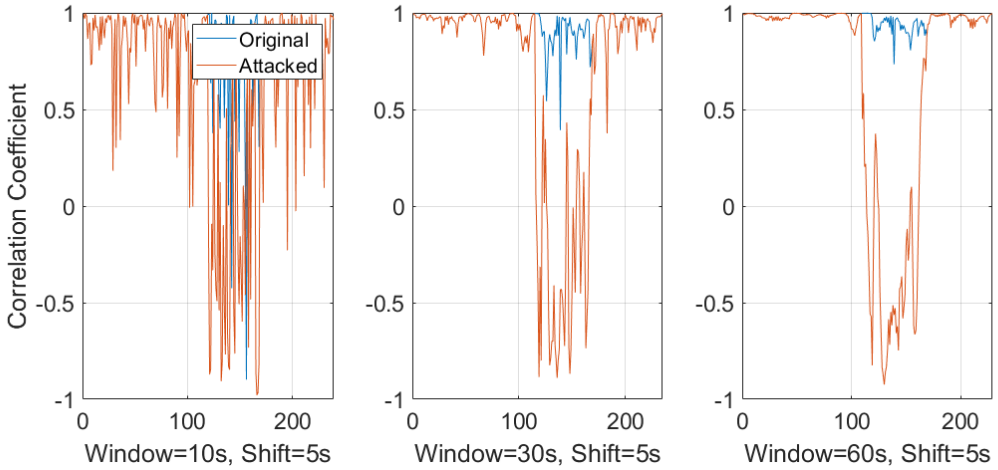


Figure 13. Correlation Coefficient with 10sec,30sec and 60sec sliding window with a shift step size of 5 seconds.

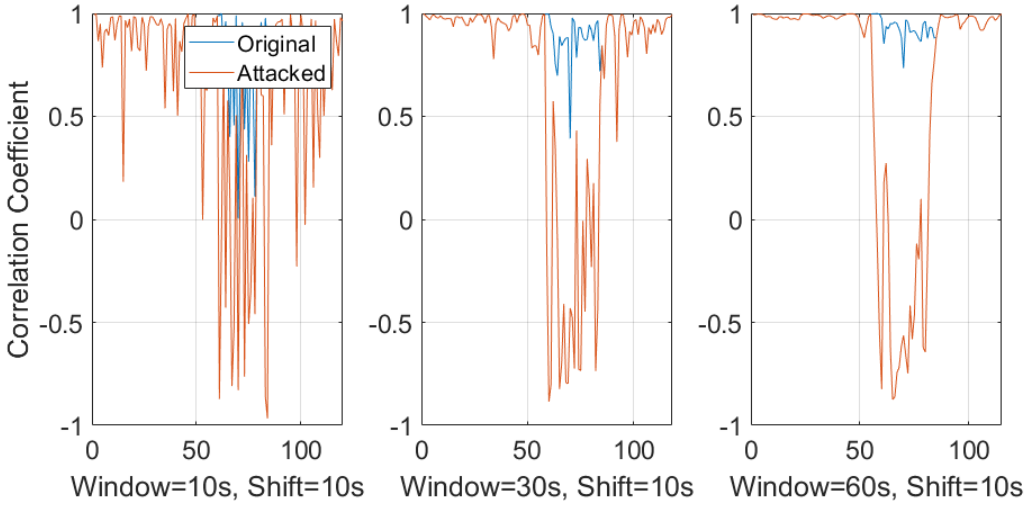


Figure 14. Correlation Coefficient with 10sec,30sec and 60sec sliding window with a shift step size of 10 seconds.

In summary, the collected data and the experimental results conclude that it is worthy to having a higher initialization setup delay with a better performance with an average shifting length of five seconds. In order to reduce the false alarm rate, a consecutive lower correlation coefficient detected by the system can be treated as an immediate alert to a challenging situation. In addition, a second pass performed by the fog layer can also be used as a reassurance for the the alert.

6. Conclusions

Increasing number of attacks on smart surveillance systems present more concerns on security. In this paper, we discussed a visual-data layer attack on video surveillance systems and introduced a novel detection method leveraging the Electrical Network Frequency (ENF) signals. ENF fluctuations are inferred to be similar at different locations at the same time instant, and these ENF traces are embedded in media recordings through various factors. The ENF estimations from the power and the

Peer-reviewed version available at Sensors 2019, 19, 2424; doi:10.3390/s19112424

audio recordings are estimated simultaneously, and a correlation coefficient is used to evaluate the signal similarity. A low correlation coefficient indicates that the signals are not similar, which in turn implies the potential existence of maliciously injected duplicated frames. A sliding window-based approach is proposed for online detection, and different parameter values are investigated to obtain a best setting.

While the proposed system is focused on audio recording to detect frame duplication attacks using ENF fluctuations at edge devices at a low computational cost, it is also possible that the ENF harmonics are contaminated due to other electromagnetic interference and affect the ENF signal estimation. To establish a secondary reliable system, our ongoing work includes developing lightweight estimating method using the ENF from the video recordings, and using the proposed technique to achieve more robust real-time authentication method for smart surveillance.

Abbreviations

The following abbreviations are used in this manuscript:

ENF	Electrical Network Frequency
FFI	False Frame Injection
STFT	Short Time Fourier Transform
FFT	Fast Fourier Transform
NFFT	Number of FFT bins
SNR	Signal to Noise Ratio
POV	Point of View
QR	Quick Response Code
AC	Alternating Current
CCD	Charge Couple device
CMOS	Complimentary Metal Oxide Semiconductor
FPS	Frames per second
HOG	Histogram of Oriented Gradients
PSD	Power Spectral Density

1. Nikouei, S.Y.; Xu, R.; Nagothu, D.; Chen, Y.; Aved, A.; Blasch, E. Real-time index authentication for event-oriented surveillance video query using blockchain. *arXiv preprint arXiv:1807.06179* **2018**.
2. Nikouei, S.Y.; Chen, Y.; Song, S.; Xu, R.; Choi, B.Y.; Faughnan, T. Smart Surveillance as an Edge Network Service: From Harr-Cascade, SVM to a Lightweight CNN. 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC). IEEE, 2018, pp. 256–265.
3. Nikouei, S.Y.; Chen, Y.; Song, S.; Xu, R.; Choi, B.Y.; Faughnan, T.R. Real-time human detection as an edge service enabled by a lightweight cnn. 2018 IEEE International Conference on Edge Computing (EDGE). IEEE, 2018, pp. 125–129.
4. Hampapur, A.; Brown, L.; Connell, J.; Pankanti, S.; Senior, A.; Tian, Y. Smart surveillance: applications, technologies and implications. Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on. IEEE, 2003, Vol. 2, pp. 1133–1138.
5. Costin, A. Security of CCTV and video surveillance systems: threats, vulnerabilities, attacks, and mitigations. Proceedings of the 6th International Workshop on Trustworthy Embedded Devices. ACM, 2016, pp. 45–54.
6. Ulutas, G.; Ustubioglu, B.; Ulutas, M.; Nabiyev, V. Frame duplication/mirroring detection method with binary features. *IET Image Processing* **2017**, *11*, 333–342.
7. Nagothu, D.; Schwell, J.; Chen, Y.; Blasch, E.; Zhu, S. A Study on Smart Online Frame Forging Attacks against Video Surveillance System. *arXiv preprint arXiv:1903.03473* **2019**.
8. Grigoras, C.; Smith, J.; Jenkins, C. Advances in ENF database configuration for forensic authentication of digital media. Audio Engineering Society Convention 131. Audio Engineering Society, 2011.

Peer-reviewed version available at Sensors 2019, 19, 2424; doi:10.3390/s19112424

9. Garg, R.; Varna, A.L.; Hajj-Ahmad, A.; Wu, M. "Seeing" ENF: power-signature-based timestamp for digital multimedia via optical sensing and signal processing. *IEEE Transactions on Information Forensics and Security* **2013**, *8*, 1417–1432.
10. Hajj-Ahmad, A.; Garg, R.; Wu, M. Spectrum combining for ENF signal estimation. *IEEE Signal Processing Letters* **2013**, *20*, 885–888.
11. Wang, W.; Farid, H. Exposing digital forgeries in video by detecting duplication. Proceedings of the 9th workshop on Multimedia & security. ACM, 2007, pp. 35–42.
12. Singh, V.K.; Pant, P.; Tripathi, R.C. Detection of frame duplication type of forgery in digital video using sub-block based features. International Conference on Digital Forensics and Cyber Crime. Springer, 2015, pp. 29–38.
13. Wahab, A.W.A.; Bagiwa, M.A.; Idris, M.Y.I.; Khan, S.; Razak, Z.; Ariffin, M.R.K. Passive video forgery detection techniques: a survey. 2014 10th International Conference on Information Assurance and Security. IEEE, 2014, pp. 29–34.
14. Fadl, S.M.; Han, Q.; Li, Q. Authentication of surveillance videos: detecting frame duplication based on residual frame. *Journal of forensic sciences* **2018**, *63*, 1099–1109.
15. Brixen, E.B. ENF; Quantification of the magnetic field. Audio Engineering Society Conference: 33rd International Conference: Audio Forensics-Theory and Practice. Audio Engineering Society, 2008.
16. Chai, J.; Liu, F.; Yuan, Z.; Conners, R.W.; Liu, Y. Source of ENF in battery-powered digital recordings. Audio Engineering Society Convention 135. Audio Engineering Society, 2013.
17. Fechner, N.; Kirchner, M. The humming hum: Background noise as a carrier of ENF artifacts in mobile device audio recordings. IT Security Incident Management & IT Forensics (IMF), 2014 Eighth International Conference on. IEEE, 2014, pp. 3–13.
18. Hajj-Ahmad, A.; Wong, C.W.; Gambino, S.; Zhu, Q.; Yu, M.; Wu, M. Factors Affecting ENF Capture in Audio. *IEEE Transactions on Information Forensics and Security* **2019**, *14*, 277–288.
19. Garg, R.; Varna, A.L.; Wu, M. Seeing ENF: natural time stamp for digital video via optical sensing and signal processing. Proceedings of the 19th ACM international conference on Multimedia. ACM, 2011, pp. 23–32.
20. Vatansever, S.; Dirik, A.E.; Memon, N. Detecting the Presence of ENF Signal in Digital Videos: A Superpixel-Based Approach. *IEEE Signal Processing Letters* **2017**, *24*, 1463–1467.
21. Bykhovsky, D.; Cohen, A. Electrical network frequency (ENF) maximum-likelihood estimation via a multitone harmonic model. *IEEE Transactions on Information Forensics and Security* **2013**, *8*, 744–753.
22. Ojowu, O.; Karlsson, J.; Li, J.; Liu, Y. ENF extraction from digital recordings using adaptive techniques and frequency tracking. *IEEE Transactions on Information Forensics and Security* **2012**, *7*, 1330–1338.
23. Rodríguez, D.P.N.; Apolinário, J.A.; Biscainho, L.W.P. Audio authenticity: Detecting ENF discontinuity with high precision phase analysis. *IEEE Transactions on Information Forensics and Security* **2010**, *5*, 534–543.
24. Su, H.; Hajj-Ahmad, A.; Garg, R.; Wu, M. Exploiting rolling shutter for ENF signal extraction from video. Image Processing (ICIP), 2014 IEEE International Conference on. Citeseer, 2014, pp. 5367–5371.
25. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. international Conference on computer vision & Pattern Recognition (CVPR'05). IEEE Computer Society, 2005, Vol. 1, pp. 886–893.
26. Zheng, Y.; Blasch, E.; Liu, Z. *Multispectral Image Fusion and Colorization*; SPIE Press, 2018.
27. Chen, N.; Chen, Y.; Song, S.; Huang, C.T.; Ye, X. Smart urban surveillance using fog computing. Edge Computing (SEC), IEEE/ACM Symposium on. IEEE, 2016, pp. 95–96.
28. Nagothu, D.; Xu, R.; Nikouei, S.Y.; Chen, Y. A microservice-enabled architecture for smart surveillance using blockchain technology. *arXiv preprint arXiv:1807.07487* **2018**.
29. Xu, R.; Nikouei, S.Y.; Chen, Y.; Polunchenko, A.; Song, S.; Deng, C.; Faughnan, T.R. Real-Time Human Objects Tracking for Smart Surveillance at the Edge. 2018 IEEE International Conference on Communications (ICC). IEEE, 2018, pp. 1–6.
30. Blasch, E. NAECON08 grand challenge entry using the belief filter in audio-video track and ID fusion. Aerospace & Electronics Conference (NAECON), Proceedings of the IEEE 2009 National. IEEE, 2009, pp. 296–303.