EAI Endorsed Transactions
on Security and Safety

Research Article EAI.EU

Authenticating Video Feeds using Electric Network Frequency Estimation at the Edge

Deeraj Nagothu[1], Yu Chen[1,*], Alexander Aved[2], Erik Blasch[2]

[1]Department of Electrical and Computer Engineering, Binghamton University, USA
[2]The U.S. Air Force Research Laboratory, USA

Abstract

Large scale Internet of Video Things (IoVT) supports situation awareness for smart cities; however, the rapid development in artificial intelligence (AI) technologies enables fake video/audio streams and doctored images to fool smart city security operators. Authenticating visual/audio feeds becomes essential for safety and security, from which an Electric Network Frequency (ENF) signal collected from the power grid is a prominent authentication mechanism. This paper proposes an ENF-based Video Authentication method using steady Superpixels (EVAS). Video superpixels group the pixels with uniform intensities and textures to eliminate the impacts from the fluctuations in the ENF estimation. An extensive experimental study validated the effectiveness of the EVAS system. Aiming at the environments with interconnected surveillance camera systems at the edge powered by an electricity grid, the proposed EVAS system achieved the design goal of detecting dissimilarities in the image sequences.

Received on 14 December 2020; accepted on 26 January 2021; published on 04 February 2021

Keywords: Video Data Authentication, Electrical Network Frequency (ENF) Estimation, Internet of Video Things (IoVT), Edge Computing, Visual Layer Backdoor Attacks

doi:10.4108/eai.4-2-2021.168648

1. Introduction

The Internet of Video Things (IoVT) has become a fundamental part of the infrastructure in smart cities, where situation awareness (SAW) plays a critical role in city monitoring and management [1], [2]. Expanding deployment of the IoVT systems leads to a vast volume of visual data captured and processed every minute. The data growth has also increased the requirements for a more scalable, flexible, and reliable IoVT system [3], [4], which is very challenging for human-in-loop platforms. Consequently, artificial intelligence (AI) based computer vision has been widely recognized as the core of next-generation IoVT [5]. A trained AI deep learning computer vision technique seeks to emulate how humans perceive the visual information [6], leading to an enhanced and secure infrastructure environment.

Compared to other embedded systems, IoVT systems have an additional level of abstraction, the visual layer, which opens a new dimension for attackers or abusers [7], [8] that should be closed. In a smart surveillance framework, the visual layer information serves as pseudo sensing for monitoring the security of the infrastructure [9], [10], [11]. By taking advantage of the imagery semantics and target recognition in modern AI powered surveillance systems, attackers can potentially mislead the operator, hide malicious activities, or get around detection algorithms. The visual layer backdoor has been identified on CCTV (closed-circuit television) cameras [12] and on a full-body scanner [13]. These visual layer backdoor attacks can be installed either locally through malicious updates over a Universal Serial Bus (USB) port or remotely via a command injection or a malicious firmware update over a web interface [14], [15], [16]. The malicious component is triggered and controlled via an unique imagery input. The trigger can be Quick Response (QR)-like codes or pre-defined imagery printed on T-shirts, cars, or any accessory visible to the cameras [17], [18].

[*]Corresponding author. Email: ychen@binghamton.edu

The lack of authentication tools with the same level of proficiency as the forging tools necessitates the development of multimedia forensic tools [19]. Using machine learning-based tools for verifying multimedia recordings could be standardized for reliable forensic fingerprint analysis. Authenticating a multimedia recording and then making decisions based on the authenticity, complements the reliability on the IoVT systems at the edge, and reduces the downtime from cloud-based applications. *Electric Network Frequency* (ENF) is one fingerprinting technique available for authentication of digital recordings due to its instantaneous behavior changing with time. ENF is a time-varying signal fluctuating across its nominal frequency 50Hz or 60Hz based on the power supply-demand from electrical power grids. Imbalance in power consumption and power flow are causes of the instantaneous variations of the ENF [20], where the fluctuations are consistent throughout the interconnect power grid. The deviation of ENF from its nominal frequency in the United States is between [-0.02, 0.02] Hz and [-0.05, 0.03) Hz in Asian and European countries.

In digital video recordings, the ENF traces occur from the light source connected to the power supply grid. As the light source flickers at both negative and positive cycles of the alternating current (AC), the illumination frequency becomes twice that of the nominal frequency [21]. Whereas in digital audio recording, ENF traces occur as a result of electromagnetic field interference by direct connection to a power grid or acoustic hum from devices connected to power grid. Estimating ENF traces from the digital recordings adds a layer of authenticity to video recordings.

For the IoVT framework, emphasizing video recordings, there are two different types of ENF estimation in digital video recordings based on the type of imaging sensor used, charge-coupled device (CCD) and complementary metal-oxide-semiconductor (CMOS). The difference between the imaging sensors used is the type of shutter mechanism implemented. For *CCD sensors*, the pixels on the sensor grid capture the visible light at the same time instant, known as the global shutter mechanism. For *CMOS sensors*, each row in the pixel grid captures the visible light sequentially at different time instants, known as the *rolling shutter* mechanism. Since the majority of the sensors in the IoVT environment consist of CMOS sensors, the proposed ENF estimation algorithm assumes that the video frames captured utilize the rolling shutter mechanism. A comparison between ENF estimations from both types of imaging sensors justifies the use of CMOS sensors for the proposed technique.

In the IoVT infrastructure, strategically deployed surveillance cameras monitor the movements of objects of interest with minimal blind spots. Processing video frames with subjects moving can cause deformation in the pixel values. For the estimation of ENF from video recordings consisting of moving subjects, an effective algorithm addresses the challenges of compensating for occlusion caused by moving subjects.

In this paper, we propose an *ENF-based Video Authentication scheme leveraging Superpixel masking* (EVAS) using the rolling shutter mechanism. The masking enables a novel ENF estimation method using Selective Superpixel Masking (SSM) and implements a non-parametric based spectrogram method in which the weighted energy is adopted to estimate the ENF. In this work, we assume that ENF traces are present in video recordings under light sources like fluorescent, incandescent, or Light Emitting Diode (LED) lights in an indoor setting, and the attacker can modify the incoming video frames by frame injection or frame duplication attacks. The paper's contributions are as follows:

- A Superpixel Segmentation algorithm is introduced that compensates for occlusion caused by moving subjects;
- A comparison of ENF estimates from video frames captured using CMOS sensors using a Rolling Shutter mechanism with and without the proposed superpixel segmentation algorithm;
- A dynamic cross-correlation coefficient is adopted that verifies the authenticity of the ENF estimate with a parallel ground truth ENF estimate from the main power grid;
- By comparing ENF estimates generated by different camera devices in a heterogeneous IoVT environment using a cross-correlation coefficient, the EVAS scheme can be further expanded to be deployed independently of the power grid module when more devices are attached; and
- A proof-of-concept prototype is built and tested using real-world scenarios, and the results verify that the EVAS scheme meets the design goals.

The rest of this paper is structured as follows. Section 2 provides background knowledge for readers along with a brief review of related work. Section 3 describes the mathematical model for ENF estimation and the key components of the EVAS scheme. Section 4 presents the experimental results and the discussions are in Section 5. Section 6 concludes the paper.

2. Background and Related Work

Electric Network Frequency (ENF), as a digital fingerprint technique, provides for forensic audio, video, and telecommunication analysis [20]. The application of ENF to authenticate multimedia recording for jurisdiction purposes has paved the way for more applications such as multimedia synchronization, geographical location tagging, detecting audio forgeries in surveillance network and verification of time of recording [7], [22], [23], [24].

Due to relatively advanced and diverse application of ENF, signal processing techniques like multi-harmonic spectral combination [25], high precision phase analysis [26], spectrogram estimation methods [20] and frequency tracking techniques [27] allows for reliable ENF estimations in signals with lower signal-to-noise ratio (SNR) augmenting ENF-based forensic applications. Forensic operations like spectrogram and inter-frequency consistency check to support the resilience of ENF against anti-forensic techniques [28]; hence, ENF is adapted as a reliable environmental fingerprint for video authentication.

2.1. ENF Estimation using Video Recordings

For IoVT environments that include audio and video, authentication adds a dual-layer security. Imaging sensors used in video cameras measure the intensity of photons falling on a sensor array and convert them to an electric current to produce a digital image. These photons collected under indoor light sources, that run on the power grid, carry ENF traces in the form of illumination frequency [21]. The current in the light source changes polarity twice that of the nominal frequency, and hence the illumination frequency is 100/120 Hz.

Researchers have tested ENF estimation in various indoor light sources and different compression ratios [29]. The experimental results yield that a minimum data compression of 500 kbps, and under LED illumination, the ENF estimation is quite robust. To confirm the presence of ENF in indoor lighting, we used a photodiode BPW21 with high spectral sensitivity in the visible range [21] to measure illumination frequency. The spectral estimation techniques like Short-time Fourier transform (STFT), followed by quadratic interpolation or weighted energy from spectral bins estimated the ENF from the recordings made in the scenario.

Figure 1 shows two optical sensor readings along with simultaneous power recordings. The first recording represents the optical sensor placed under a LED light source. For the second recording, the sensor recorded the ambient light in the room in the absence of a direct overhead light source. It is clear that the ENF signal is present in the light source, and the captured video recordings under these light sources contain ENF traces. The correlation coefficient between the two recordings determines the similarity of frequency fluctuations. In Fig. 1, the optical sensor recorded light fluctuations at 120Hz; and for convenient representation, the illumination frequency and collected power ENF are compared at 60Hz.

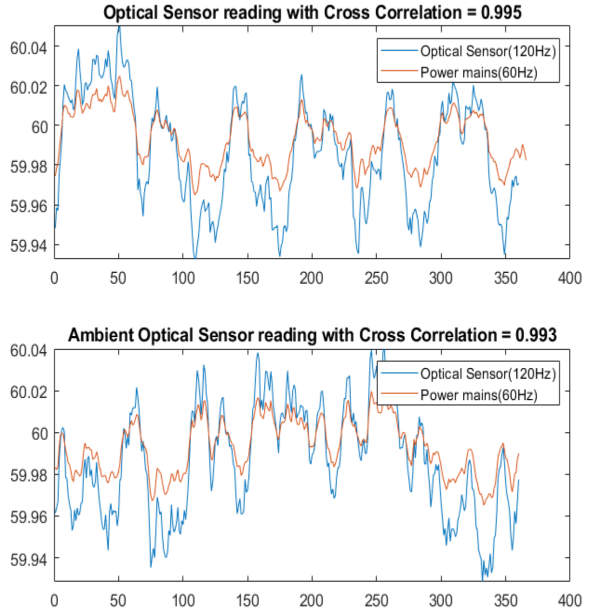

Figure 1. Optical sensor reading collected under direct mains powered light source and ambient light. Compared with simultaneous power recording using Cross Correlation Coefficient.

Sensors in cameras like CCD or CMOS capture visual data at the rate of 25/30 frames per second (FPS). The nominal frequencies of ENF are 50/60 Hz in different parts of the world. A lower sampling rate of cameras with 25/30 FPS introduces significant aliasing of the ENF component in video recordings due to the Nyquist criterion. The majority of the video cameras are not truly 25/30 FPS; instead, they capture at 23.98/29.97 FPS due to the video standards established. The aliasing effect causes the 25/30 FPS sampling to disappear as the DC component, whereas 23.98/29.97 FPS causes the ENF to appear at different aliasing frequencies, determined from the sampling theorem

$$f_a = |f_l - k \cdot f_v| < \frac{f_v}{2} \tag{1}$$

where f_a is the aliasing frequency, f_l is the illumination frequency, f_v is the sampling rate of video recorders, i.e. video FPS, and k varies until the condition is satisfied. Table 1 presents different f_a for their respective nominal frequency.

Based on the types of imaging sensors, there are different ENF extraction techniques. In the case of CCD sensors, all the exposed imaging pixels on the sensor capture the visible photons at the same time, also known as *global shutter* sensors. Sampling the mean of pixel intensities in frames gives video samples. Using the aliasing frequency, the ENF is estimated. For

Table 1. Aliased Frequency for a given Video Frame rate with different nominal frequency.

Nominal Frequency (Hz)	Frame Rate	Aliasing Frequency (Hz)	2nd Harmonic
60	29.97	0.12	0.24
60	30	0	0
60	25	5	10
50	29.97	10.09	9.79
50	25	0	0
50	30	10	10

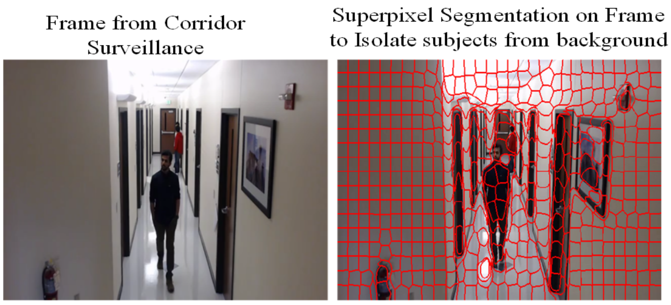
Frame from Corridor Surveillance
Superpixel Segmentation on Frame to Isolate subjects from background

Figure 2. Superpixel segmentation on a surveillance frame.

example, a 720p video recording of 5 minutes in CCD with 23.987 FPS has an aliasing frequency of 0.12Hz, where the nominal frequency is 60Hz from Eq. (1), and the number of samples obtained are 7196 samples.

In the case of the CMOS sensor, a sequential exposure of the sensor pixels to light results in different rows in the pixel grid having different light exposures at unique time instances. The shutter mechanism in which the sensor and imagery pixels are sequentially exposed to light is also called a *rolling shutter mechanism*. Each row represents one sample of pixel intensities followed by an idle period before the next frame, which collectively increases the temporal resolution of video samples satisfying the Nyquist criterion [30]. For example, a 720p video recording of 5 minutes in CMOS with 23.987 FPS, from Eq. (2) the number of samples obtained are 5,181,192 samples, which is comparatively very high from CCD sensors, as by:

$$F_{s_{(CMOS)}} = Frame_{height} * Video_{FPS} \tag{2}$$

where $F_{s_{(CMOS)}}$ is the sampling frequency of CMOS sensor video recording. With a better sampling frequency, the ENF estimation from CMOS sensors yields better results compared to the CCD sensors. Recent advancements in ENF estimation involves efficient idle period estimation [31], [32], since the idle period results in pixels with no exposure, i.e., missing ENF samples. For applications in surveillance recordings using a rolling shutter mechanism, each sample is represented by each row in a frame. The videos include moving subjects, which causes non-rigid deformation or occlusion. To reserve the computational load by looking for ENF artifacts in all video recordings, a superpixel based approach to verify the presence of ENF has been introduced. Authors in [33] verified the presence of ENF in global shutter mechanism based video recordings using superpixels. This technique only validated the presence of ENF, in a video recording by comparing different superpixels.

For indoor video recordings, in the IoVT infrastructure, the chances of capturing ENF traces are much higher through indoor lighting. The ENF traces from the video feed compared with ENF collected from the power grid for the same time instant can serve as a standard video authentication system. A vast deployment of the CMOS sensors in mobile devices, portable computers, and the surveillance cameras, which altogether represents the majority of the devices in the IoVT infrastructure, provides enough basis to focus on the CMOS sensors. This paper explores the presence of ENF traces from CMOS sensors in IoVT recordings and tackles the occlusion problem caused due to the movement of the subjects in the video frames using superpixels. Compared to [33], we adopt superpixels to tackle the occlusion problem due to moving subjects in the rolling shutter based video recordings.

2.2. Superpixels Application in Video Recordings

Segmentation of image involves grouping the pixels with similar intensity and texture pattern, which divides the image into non-overlapping sub-pixels known as superpixels. Instead of pixel-wise computation, superpixel computation has decreased the image-based computational load. Grouping of pixels with similar spatial features is used for applications like edge-detection, classification, and recognition. In this work, a gradient-ascent based algorithm with k-means clustering is adopted [34], also known as the *Simple Linear Iterative Clustering (SLIC)* algorithm. The SLIC algorithm was preferred over other graph-based algorithms due to better memory efficiency, segmentation performance, and fast computation. Figure 2 represents the superpixel segmentation of a frame collected from indoor corridor surveillance. Here, the moving subject can be distinguished from the static background using superpixels.

An earlier tracking algorithm separates the moving subject from its background with superpixels using mid-level cues [35], which handles heavy occlusion and shows that superpixel segmentation for motion tracking yielding better performance. A robust background initialization algorithm using superpixels was introduced recently [36], in which the stored background from the sub-sequences removes the moving subjects from the frame and generates reliable background candidates.

3. EVAS: Mathematical Model and Detection

Calculating pixel intensities for ENF estimation faces challenges like moving subjects in the frame, which cause undesired changes in intensity values. For a better ENF estimate from these video recordings, addressing the dynamic nature of this problem is required. In the EVAS scheme, we introduce a *Selective Superpixel Masking (SSM)* technique to address the challenges of moving objects in the image.

3.1. Fundamentals of Superpixel Segmentation

Video frames collected in the IoVT environment have a static background recorded with a stationary camera for most infrastructures. Based on this assumption, the discrepancies caused by moving subjects are separated from the stationary frame using superpixels. Based on the SLIC algorithm [34], cluster centers are initialized by assuming N pixels in the image. The number of pixels in each superpixel is $\frac{N}{K}$, where K is the parameter controlling superpixel size. For the CIELAB color space representation, the center of each superpixel is initialized as

$$C_i = [l_i, a_i, b_i, x_i, y_i]^T_{i=1,...,K} \tag{3}$$

where C_i is the i^{th} cluster center, $[l_i, a_i, b_i]$ are components of lab color space, and $[x_i, y_i]$ are the pixel position.

The *lab* color space values vary in a known range, whereas the pixel position values vary based on the frame resolution. The distance D between the i^{th} pixel and the K^{th} cluster center estimates whether the pixel belongs to that superpixel cluster. The distance measure is calculated as

$$D = \sqrt{{d_c}^2 + \frac{d_s}{S}^2 m^2} \tag{4}$$

$$d_c = \sqrt{(l_k - l_i)^2 + (a_k + a_i)^2 + (b_k + b_i)^2} \tag{5}$$

$$d_s = \sqrt{(x_k - x_i)^2 + (y_k - y_i)^2} \tag{6}$$

where d_c and d_s are distance measurements of color proximity and spatial proximity. S is the approximate superpixel grid size $(S \times S)$ and is given as $S = \sqrt{\frac{N}{K}}$. Lastly, m is the weighting factor between color and spatial difference ranging from $[1, 40]$, where a larger value enforces a superpixel with more regular and smoother shapes. In this study, we selected a m value of 10 after several experimental tests.

For a k-means algorithm, the computation complexity of superpixel segmentation is $O(N)$, where N is the number of pixels. The SLIC algorithm is both computationally and memory efficient, and the advantages of the algorithm increases with the size of the frame.

3.2. Selective Superpixel Masking

Moving subjects in the frame are a common problem for analyzing video frames using pixel intensities, due to non-rigid deformation and occlusion in uniform pixel values. The EVAS scheme proposes a selective superpixel masking (SSM) algorithm to compensate for the motion detected in the video. SSM uses frame segmentation on consecutive video frames and compares the superpixel similarity among these frames. Any inconsistencies in the pixel values are masked, leading to more uniform pixel values.

For a given video frame sequence $F = \{F_n\}_{n=1,...,M}$, where M is the total number of frames in the given sequence, a Gaussian mixture model (GMM) is used. The GMM is a simple non-parametric adaptive density estimation method for background subtraction [37], which generates the motion mask $D_{x,y}$. The background for the test videos are largely static and unchanging over consecutive frames of a video. Therefore, with a GMM model, the foreground is segmented from the background allowing any substantial change like moving subjects. The obtained matrix $D_{x,y}$ consists of the subject motion in the form of a logical matrix, and it is compared with the superpixel segmentation $\mathcal{S}_n$ of the frame.

$$\mathcal{S}_n^{'} = \mathcal{S}_n . * D_{x,y} \tag{7}$$

$$\mathcal{M}_n^{'} = \mathcal{S}_n - \mathcal{S}_n^{'} \tag{8}$$

Here $\mathcal{S}_n^{'}$ is a superpixel frame which carries the individual affected pixels from $D_{x,y}$ due to subject motion. By comparing $\mathcal{S}_n^{'}$ with $\mathcal{S}_n$, EVAS generates a motion mask of superpixels $\mathcal{M}_n^{'}$, which preserves the steady superpixel regions and focuses on the superpixel regions with modified pixels.

From our observations, some pixels are modified due to reflective property of objects in the frame, compared to moving subjects where the changes are drastic. So, a superpixel region is masked out when the number of pixels it contains are significantly modified by comparing it to a threshold. For all the superpixels $\mathcal{SP}_K$ in $\mathcal{M}_n^{'}$, where K is the number of superpixels in a frame, the Superpixel based Motion Mask $\mathcal{M}_n$ for moving object is given as,

$$\mathcal{M}_n = \begin{cases} \mathcal{SP}_k = 1, & \mathcal{N}(\mathcal{M}^{'}_{SP_k}) < \tau_{pixels} \\ \mathcal{SP}_k = 0, & \mathcal{N}(\mathcal{M}^{'}_{SP_k}) \geq \tau_{pixels} \end{cases} \tag{9}$$

where $\mathcal{N}(\mathcal{M}^{'}_{SP_k})$ is the number of pixels affected in a superpixel $(\mathcal{SP}_k)$ for $\mathcal{M}_n^{'}$. By comparing with a threshold τ_{pixels}, the algorithm decides if a pixel was affected due to moving subject or small environmental interference. The mask $\mathcal{M}_n$ eliminates any motion detected, and the masked superpixels are not accounted for pixel intensity vector. The irregularities like

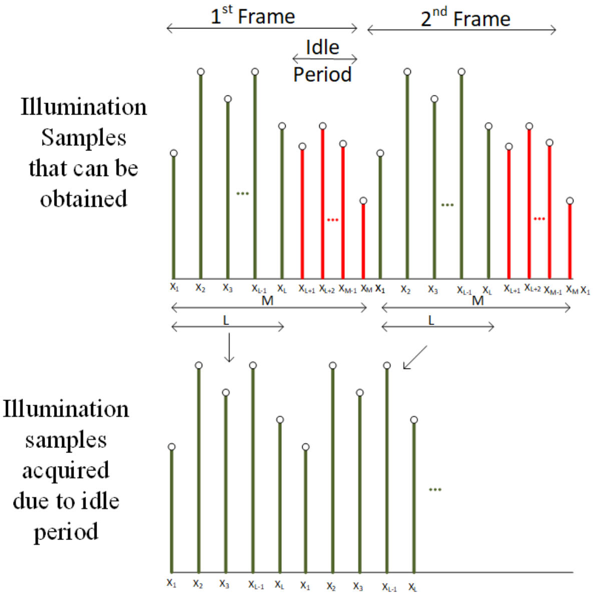
1st Frame
2nd Frame
Idle Period
Illumination Samples that can be obtained
x_1 x_2 x_3 x_{L-1} x_L x_{L+1} x_{L+2} x_{M-1} x_M
M
L
Illumination samples acquired due to idle period

Figure 3. Rolling Shutter Sampling Mechanism.

occlusion caused due to subject motion are eliminated. With the stable pixel values collected using the superpixel mask, the ENF is estimated from the steady pixel intensities obtained.

3.3. ENF Estimation Using Rolling Shutter Mechanism

With Rolling Shutter, each frame consists of multiple ENF samples where the samples are distributed sequentially from the top row of the frame to the bottom row. Due to sequential sampling, the temporal sampling rate of the video recording can be much higher compared to the ENF estimation using aliasing frequency, Eq. (2). Figure 3 represents the image acquisition mechanism in a CMOS sensor camera where the frames are sequentially exposed. Figure 4 depicts a time-domain illustration of the L-branch filter bank model. Assuming that the camera can produce M samples i.e., the number of rows per frame, the camera retains only L samples due to the idle period introduced by camera manufactures ($L \leq M$).

From the time-domain illustration, the input signal $x(n)$ represents the illumination samples when there is no idle period, and $y(n)$ represents the illumination samples with idle period i.e., after dropping some samples. For a frequency domain representation, an L-branch filter bank model is used. Here, $x(n)$ to the model is shifted back in time, followed by an M-fold down-sampling filter. Then an L-fold up-sampling filter is applied, followed by shifting the signal forward in time, resulting in the output signal $y(n)$. The discrete-time Fourier transform (DTFT) of the l^{th} branch, the frequency domain representation [32] of the output signal $Y_l(e^{j\omega})$ is represented as,

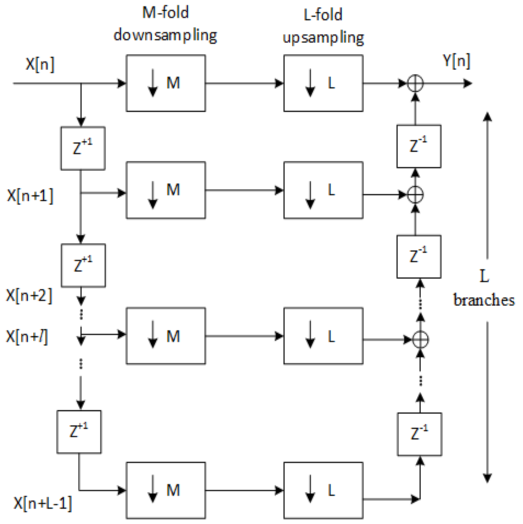

Figure 4. Filter Bank Model where ($L \leq M$) [32]

$$Y_l(e^{j\omega}) = \frac{1}{M}\left(\sum_{m=0}^{M-1} X\left(\frac{\omega L - 2\pi m}{M}\right) e^{j\frac{\omega L - 2\pi m}{M}l}\right) e^{-j\omega l}$$

where m and l varies over the actual number of row samples per frame (M) and row samples retained (L) respectively, and ω is a frequency variable with *radians/sample* unit and 2π periodicity. After combining the individual branch output $Y_l(e^{j\omega})$, and representing $Y(e^{j\omega})$ as $Y(\omega)$ for simplicity, the resulting output signal,

$$Y(\omega) = \sum_{l=0}^{L-1} Y_l(e^{j\omega})$$

$$Y(\omega) = \sum_{l=0}^{L-1} \frac{1}{M}\left(\sum_{m=0}^{M-1} X\left(\frac{\omega L - 2\pi m}{M}\right) e^{j\frac{\omega L - 2\pi m}{M}l}\right) e^{-j\omega l}$$

$$Y(\omega) = \sum_{m=0}^{M-1} X\left(\frac{\omega L - 2\pi m}{M}\right) F_m(\omega) \tag{10}$$

where

$$F_m(\omega) = \frac{1}{M}\sum_{l=0}^{L-1} e^{-j\frac{\omega(M-L)+2\pi m}{M}l}$$

In Eq. (10), the frequency-domain representation shows how the input visual signal through pixel intensities is affected due to the camera image acquisition system. The attenuation in the ENF signal is represented using F_m, depending on the proportions of L to M. The idle period specific to individual camera manufacturer can be estimated [32] by finding the emerging shifted illumination frequency using Eq. (10).

The ENF signal fluctuations are embedded in these video recordings in the form of illumination frequency, along with the steady video content. For a static video with no object movement, the row signal $R(r,n)$ can be represented as a sum of video content $V(r,n)$ and ENF signal $E(r,n)$. Here r is the row position and n is the frame number.

$$R(r,n) = V(r,n) + E(r,n)$$

Evaluating the average value of each of the signals, we can see that for a static video $V(r,n)$ is constant whereas the average of $E(r,n)$ is 0 for a large number of frames since its value fluctuates around the nominal value. Subtracting the averages from the row signal,

$$\hat{R}(r,n) = R(r,n) - \bar{R}(r)$$
$$\hat{R}(r,n) = R(r,n) - V(r)$$
$$\hat{R}(r,n) = E(r,n)$$

ENF signals can be estimated from a static video by steady-content analysis. For video with motion, a real-time Superpixel Segmentation Mask (SSM) from Eq. 9 is applied to ignore the moving subjects resulting in video samples collected from static background. Using the non-parametric spectrogram estimation methods, the ENF is estimated from the evaluated row signal.

$$\tilde{R}(r,n) = R(r,n) \odot \mathcal{M}_n$$

3.4. Measure of Similarity: Correlation Coefficient

ENF estimation from two recordings are compared based on the Pearson Correlation Coefficient metric (ρ). The ENF signal from power P_{ENF} and video V_{ENF} is given as

$$\rho(l) = \frac{\sum_{t=1}^{N}[f_{P_{ENF}}(t) - \mu_{P_{ENF}}][f_{V_{ENF}}(t-l) - \mu_{V_{ENF}}]}{var(P_{ENF}) * var(V_{ENF})} \quad (11)$$

where $f_{P_{ENF}}$ and $f_{V_{ENF}}$ are the ENF frequency estimation from simultaneous power and video recordings. l is the lag between the two signals, μ is the mean of the signal, and var is the variance of the signal. In the next section, we report the experimental studies based on the proposed model of ENF estimation using the selective superpixel masking.

4. Experimental Study

For a reliable ENF estimation from video recordings, the pixel intensities should be free from any deformation caused by a moving subject in the frame. As discussed in Section 3.2, masking the moving subject from the frame enables a more reliable and uniform pixel intensity extraction. The SLIC algorithm is used for segmentation of the frames in a video sub-sequence, and based on the segmentation; the SSM algorithm eliminates the pixels affected by moving subjects.

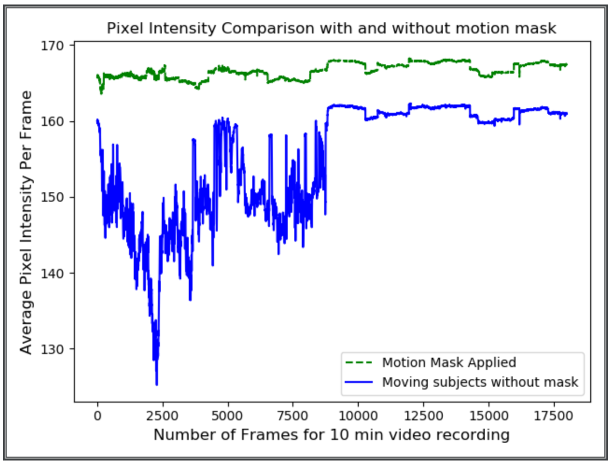

Figure 5. Comparing pixel fluctuations caused by a moving subject. A 10 minutes video recording is used where the first 5 minutes has moving subject in frame and the next 5 minutes has a static scene

4.1. Validation of the SSM Algorithm

The analysis of the video sub-sequence is first carried out to determine any moving subjects with a threshold of frame difference. Based on the identification of the frame difference, a motion mask is applied to the frame. The algorithm does not corrupt any underlying pixel data. The main objection of SSM algorithm is to extract pixel intensities from parts of image which remain unaffected by a moving subject.

Figure 5 represents the measure of impact a moving subject has on pixel intensities. A video recording of 10 minutes duration is used, where the first five minutes have a moving subject in frame and the second half includes a static scene. Average pixel intensity from each frame is used to study the effects of a moving subject. From the Fig. 5, it is clear that the impact of motion is higher and can also have a significant effect on ENF traces (further discussed in Section 4.4). By applying the proposed motion mask, the pixel fluctuation artifacts due to motion are removed in the first half of recording and the resulting pixel intensity is more stable.

The SSM algorithm is applied to different scenarios in Fig. 6. The differences in pixel intensities are first compared with the superpixel segmentation to recognize the pixel changes. In some cases, the subject in the frame is either moving slowly or stationary, which implies a more subtle inconsistency. The frame difference $D_{x,y}$ in Fig. 6 shows that the subject is not entirely covered based on the pixel difference.

Superpixels of the original frame are compared with the frame difference matrix. All the superpixels which intersect with the frame difference are masked. For the moving subject, the motion masks M_n also covers the reflection on the ground due to the lower threshold of

Δ. This superpixel based selective motion mask allows eliminating the causes of the fluctuations due to motion, and allows for a better static scene analysis in the IoVT environments.

With the superpixel masking algorithm, any detected change in the algorithm masks the entire superpixel instead of restricting it to the region of movement. The number of superpixels per-frame controls the area under each superpixel region. Regardless of the speed of the moving subject, the changes in each superpixel are masked with similar computation requirements. The sensitivity of masking can be increased with the lower number of superpixels per frame.

4.2. ENF Estimation using Rolling Shutter Mechanism

To verify the presence of ENF in a static video recording, we opted to first check the presence of ENF in a corridor surveillance video. It is clear from Fig. 7 that ENF traces are present in an indoor illuminated video recordings, in reference to the ground truth power ENF. The corridor surveillance recording was made at 30 FPS, and the ENF traces are estimated without any problem compared to the aliasing frequency technique where the ENF would be found at 0 Hz from Table 1.

4.3. Measure of Similarity

To validate the estimated ENF from both video recordings and the power recordings, a correlation coefficient is used. The value of correlation varies from [-1,1] where 1 implies highest similarity. The correlation for ENF estimations from the corridor surveillance in Fig. 7 is shown in Fig. 8. To verify the authenticity, a threshold of 0.8 is used after many observations from multiple recordings.

4.4. Affects on ENF Fluctuations with SSM

With SSM, the unnecessary fluctuations in the pixel intensities are eliminated. The resulting frame includes of steady background, from which the pixel values for each row can be extracted. Figure 9 represent the difference in the ENF estimated from the video recording with and without SSM applied. Figure 9 shows that with the proposed SSM framework, the performance of ENF estimation is improved compared to that of earlier proposed models. Since the position of camera is assumed to be stable in an indoor surveillance network, the superpixel segmentation of each frame is not necessarily evaluated. The segmentation from one frame can be applied to consecutive frames, reducing the computational complexity to compute superpixel segments per frame. Superpixel segments are periodically calculated to avoid computation and increase the pixel intensity evaluations.

The mismatch in ENF is observed in the correlation coefficient as well. The video recording includes of moving subject for first half period and then static background for rest of the recording. Figure 10 demonstrates the different in the correlation coefficient of ENF estimated from video recordings without SSM and with SSM from Fig. 9. It is clear that with a moving subject, the correlation drops significantly and could result in false negative detection. The proposed *SSM algorithm avoids the affected pixels at real-time and continues to generate reliable ENF estimations*. The SSM model compared to the earlier models is robust to the environmental noise. A minor drop in the correlation with SSM applied could potentially be due to comparison between different harmonics, or a significant number of pixels affected due to motion. For such cases, the threshold can be revised based on the deployed surveillance infrastructure.

4.5. Adaptation of proposed model in IoVT environment

It is a concern that anti-forensic tools may become capable of producing forgeries of any digital recordings [8]. A real-time implementation of such forgery attacks in the IoVT environment can be fatal for public security. Integrating the proposed authentication technique in IoVT environment reduces the detection time in case of any forgery attacks. Edge-based devices like Raspberry Pi are capable enough to handle multiple threaded processes as well as provide sufficient computational power. The EVAS system involves authenticating video feeds by comparing two simultaneous ENF, one from targeted video recording and another from ground truth power ENF. The testbed setup includes a Raspberry Pi 4 Edge based computer with a Camera attached recording at a video resolution of 1080p. For faster processing, the resolution is downsized to 720p, without any significant loss of ENF estimation. A power module with a voltage divider circuit and step down transformer is attached through USB for ground truth ENF. The Raspberry Pi can simultaneously estimate ENF from video recording and power recording using parallel threading.

To make real-time edge-based detection, the live video feed is batched into windows and then incremented in step sizes. For initialization of the system, a delay of 45-60 seconds is required to compute the ENF from first window, and then each window is shifted by 10 seconds to compute the next set of ENF correlations. Using a sliding window approach allows an online edge-based detection of any video feed tampering. The wait period allows a systematic cool down period for Raspberry Pi avoids bottle-necking problems. A comparison of multiple window sizes and shift sizes are used to compare the video ENF and power ENF in Fig.

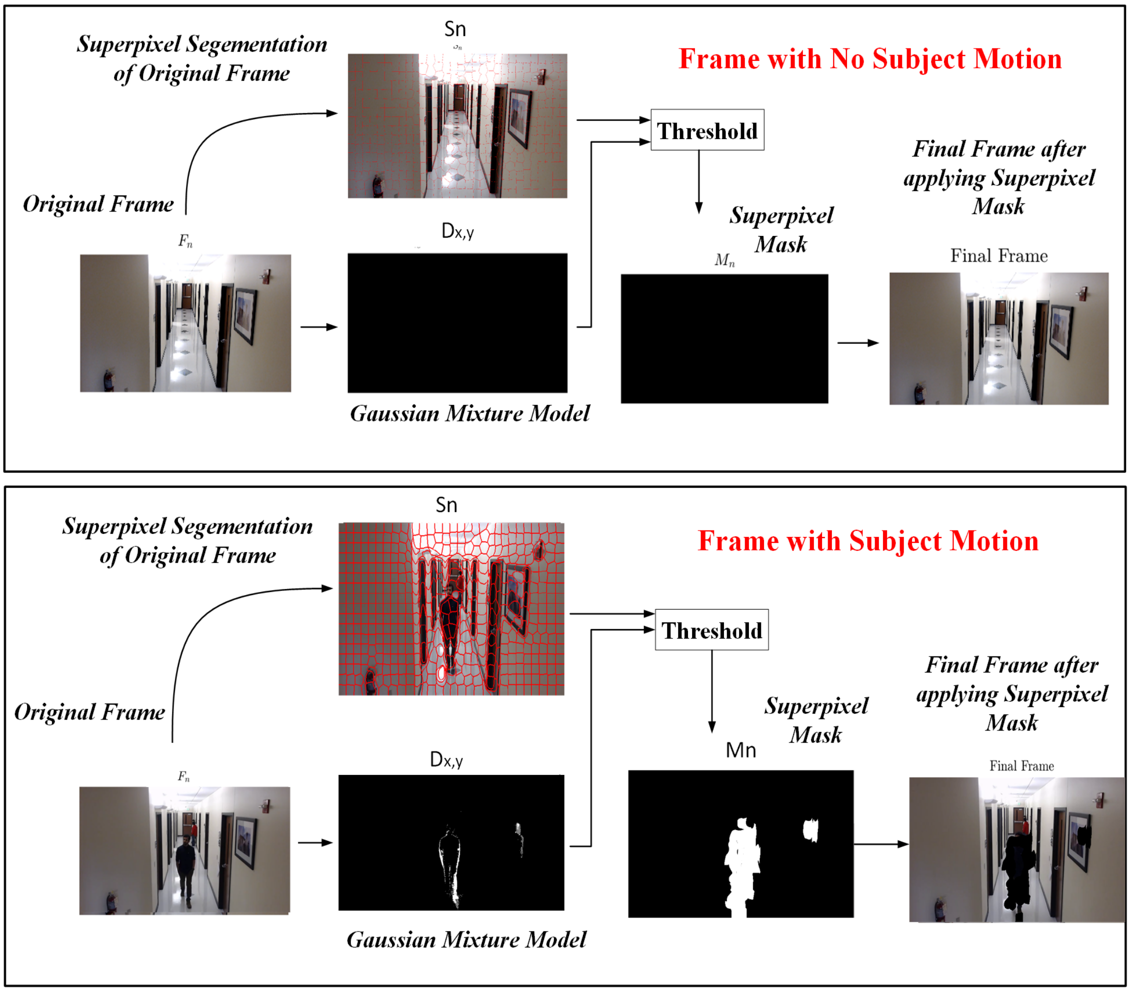
Superpixel Segementation of Original Frame
Sn
Frame with No Subject Motion
Threshold
Final Frame after applying Superpixel Mask
Original Frame
Superpixel Mask
F_n
Dx,y
M_n
Final Frame
Gaussian Mixture Model
Superpixel Segementation of Original Frame
Sn
Frame with Subject Motion
Threshold
Final Frame after applying Superpixel Mask
Original Frame
Superpixel Mask
Mn
F_n
Dx,y
Final Frame
Gaussian Mixture Model

Figure 6. Algorithm flow of the Selective Superpixel Segmentation Masking.

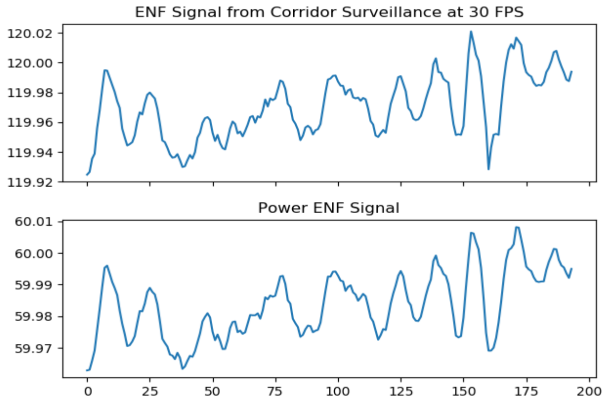

Figure 7. Comparing estimated ENF from Corridor Video Surveillance recording and parallel Power ENF recording.

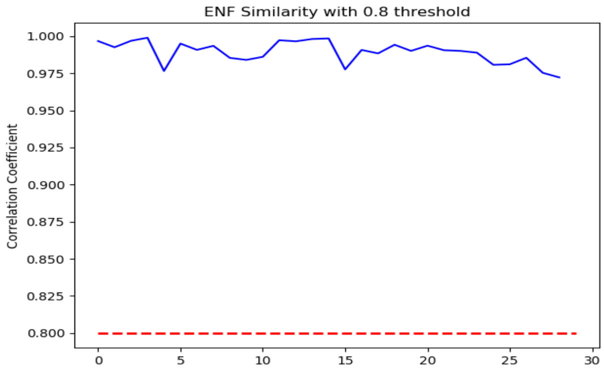

Figure 8. Correlation coefficient for Fig. 7.

11. With these observations, we used a window size of 60 seconds and shift size of 10 seconds. The shift size can further be decreased to five seconds at no cost of performance, but 10 seconds is used to avoid CPU (central processing unit) over-usage.

A *Replay attack* was performed on the video feed where the frames are recorded and continuously repeated to camouflage live events in the video. Using the EVAS system, Fig. 12 shows the mismatch in the

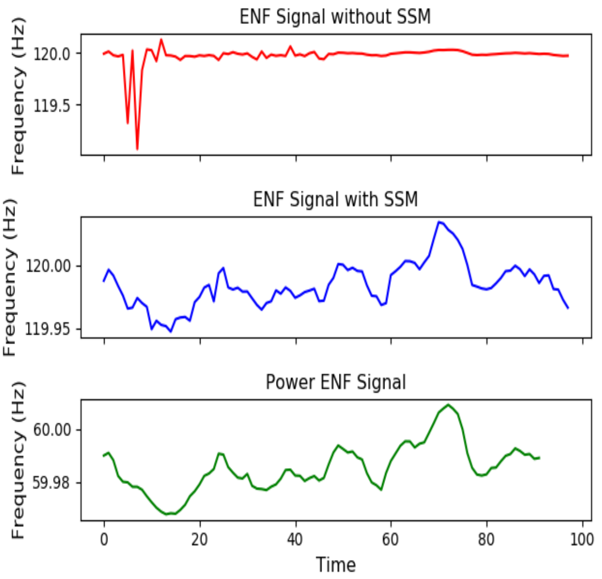

Figure 9. ENF signal of a video recording where the first half of the recording includes moving subject.

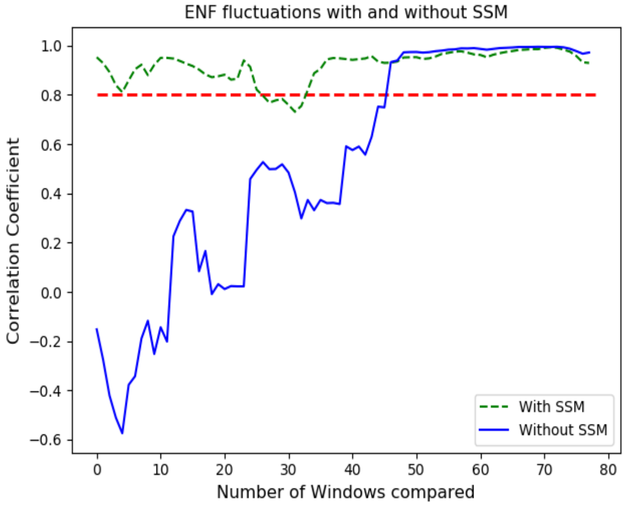

Figure 10. Correlation of Video ENF and Power ENF with and without SSM applied.

ENF signal estimated from the video recording and a significant drop in the correlation coefficient. A threshold of 0.8 is used depending on the computation complexity and the ENF signal estimation accuracy. The computation burden for STFT algorithm to estimate the spectrogram of required nominal frequency strip is,

$$\frac{l}{w-o} * NFFT * log_2(NFFT)$$

The ENF signal resolution varies based on the length of the signal used for ENF estimation (l), window size (w), overlapping window size (o), and the frequency resolution ($NFFT$) used for signal estimation. Depending on the sensitivity of the location which is monitored, the threshold requirements are adjusted to minimize false negatives. Note that modern large infrastructures include hundreds of cameras for surveillance, and continuous monitoring of such network is complicated. The EVAS system enables edge-based detection for any video tampering and notifying the surveillance authority.

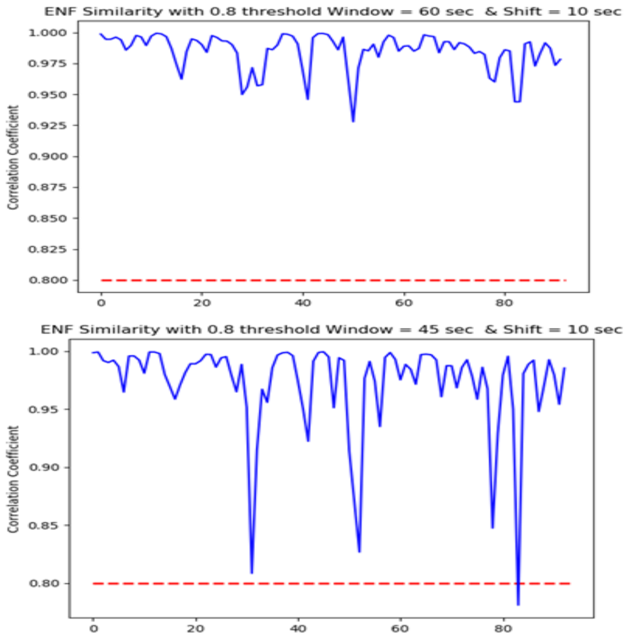

Figure 11. Comparing multiple window sizes and shift sizes for ENF comparison.

4.6. Distributed ENF estimation from multiple cameras

The surveillance camera network deployed in an indoor environment can be adapted to the proposed model by inter-authenticating the ENF signature. The nature of the ENF signal is such a way that it is similar at one time instant throughout the grid, and for the targeted framework of surveillance network the ENF should be similar for multiple cameras. To test the ENF consistency, different cameras with different frame rates were used to record a video at one time instant. Figure 13 shows that the ENFs are similar at one time instant throughout the surveillance infrastructure; which helps in reducing the redundant power module for authentication, and a distributed framework can be established to authenticate video stream on the existing system.

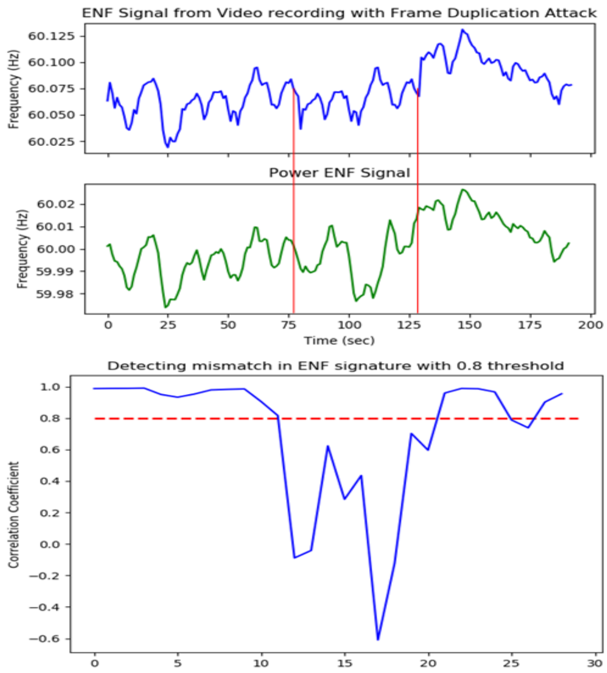

Figure 12. Detection of Video Forgery Attacks on the Surveillance system using the proposed EVAS model. The correlation coefficient is compared to the threshold 0.8.

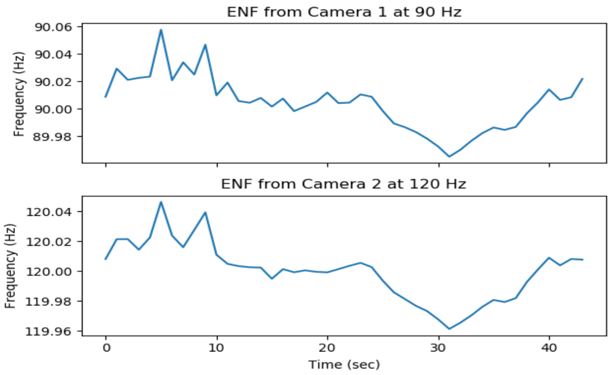

Figure 13. ENF estimated from Multiple cameras at one time instant. Camera 1 had a frame rate of 30 FPS, and camera 2 had a frame rate of 29.98 FPS.

5. Discussions

EVAS is a video authentication scheme based on an ENF estimation method, which uses the rolling shutter mechanism on extracted pixel intensities captured from video recordings in the IoVT environment. The recordings from a stationary surveillance camera might include moving subjects, which results in the unwanted pixel intensity fluctuations.

Some challenges still exist in the use of the superpixel masking by comparing consecutive frames. For example, any stationary subject might not be masked when the number of superpixels used is higher. By decreasing the number of segments per frame, any changes in the frame by movement can be captured by the whole superpixel. The pixel intensities are also affected in the situation when the camera adjusts its focus for tracking the subject, hence resulting in unwanted fluctuations. These fluctuations are caused due to a lower dynamic range of the commercial surveillance cameras. But this is an unusual problem since most of the deployed cameras include lower aperture for focusing on large region, hence reducing the occurrence of camera auto-focus.

A moving surveillance camera is also another difficult case of generating static background. To tackle the camera motion problem, a new modality for video synchronization can be adopted where shifting the pixels in two adjacent frames would produce a stable static frame [38]. This would also require high computation power along with the proposed EVAS system.

For each camera recording, the nominal frequency of the ENF in video recording depends on the frame rate used by the camera. Since the focus of this paper is in large scale deployed surveillance cameras in an indoor environment, all the cameras have similar configuration. With an initialization process and dynamic estimation of nominal frequency collected from peak frequency spectrum, the proposed model can be adapted to large scale surveillance network.

From earlier observations, eliminating the power module and depending on the existing video recording stream from multiple cameras, can help with adapting the EVAS framework with any surveillance system. As a part of our on-going efforts, we are adapting the EVAS system with a distributed system model for a cost-effective and ready-to deploy system.

6. Conclusions

The paper proposes the ENF-based Video Authentication method leveraging steady Superpixels (EVAS) framework to tackle the challenge of online video feeds authentication in an edge IoVT environment using ENF matching. EVAS verifies the authenticity of a video by reference matching the ground truth ENF with the estimated ENF from videos. From our observations, the video recordings from surveillance cameras include moving subjects, which disrupt the illumination samples resulting in inaccurate ENF estimation. Selective Superpixel Masking in the EVAS model solves the challenges of compensating for occlusion caused by moving subjects. The proposed EVAS is deployed on an edge-based system for indoor surveillance monitoring and authentication using ground truth ENF. A sliding window protocol type mechanism is used for batch verification of recorded frames.

For future directions of this work include applying the EVAS framework to more online video services like online conferences and social media video recordings. Both audio and video ENF can be synchronized for robust authentication and detecting any spatio-temporal visual layer attacks. Currently, a crawler network is being developed for verifying the authenticity of online social media videos with ENF traces as an environmental fingerprint to combat digital video forgeries.

Acknowledgement. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U. S. Air Force.

References

[1] Liu, B., Chen, Y., Shen, D., Chen, G., Pham, K., Blasch, E. and Rubin, B. (2014) An adaptive process-based cloud infrastructure for space situational awareness applications. In *Sensors and Systems for Space Applications VII* (International Society for Optics and Photonics), **9085**: 90850M.

[2] Wu, R., Liu, B., Chen, Y., Blasch, E., Ling, H. and Chen, G. (2017) A container-based elastic cloud architecture for pseudo real-time exploitation of wide area motion imagery (wami) stream. *Journal of Signal Processing Systems* **88**(2): 219–231.

[3] Gebre-Amlak, H., Lee, S., Jabbari, A.M., Chen, Y., Choi, B.Y., Huang, C.T. and Song, S. (2017) Mist: Mobility-inspired software-defined fog system. In *2017 IEEE International Conference on Consumer Electronics (ICCE)* (IEEE): 94–99.

[4] Nikouei, S.Y., Xu, R., Nagothu, D., Chen, Y., Aved, A. and Blasch, E. (2018) Real-time index authentication for event-oriented surveillance video query using blockchain. In *2018 IEEE International Smart Cities Conference (ISC2)* (IEEE): 1–8.

[5] Zhang, J. and Tao, D. (2020) Empowering things with intelligence: A survey of the progress, challenges, and opportunities in artificial intelligence of things. *IEEE Internet of Things Journal* .

[6] Nikouei, S.Y., Chen, Y., Aved, A. and Blasch, E. (2018) Toward an event-oriented indexable and queryable intelligent surveillance system. *IEEE COMSOC MMTC Communications – Frontiers* **13**(4): 33–39.

[7] Nagothu, D., Chen, Y., Blasch, E., Aved, A. and Zhu, S. (2019) Detecting malicious false frame injection attacks on surveillance systems at the edge using electrical network frequency signals. *Sensors* **19**(11): 2424.

[8] Nagothu, D., Schwell, J., Chen, Y., Blasch, E. and Zhu, S. (2019) A study on smart online frame forging attacks against video surveillance system. In *Sensors and Systems for Space Applications XII* (International Society for Optics and Photonics), **11017**: 110170L.

[9] Chen, N. and Chen, Y. (2018) Smart city surveillance at the network edge in the era of iot: opportunities and challenges. In *Smart Cities* (Springer), 153–176.

[10] Xu, R., Nikouei, S.Y., Nagothu, D., Fitwi, A. and Chen, Y. (2020) Blendsps: A blockchain-enabled decentralized smart public safety system. *Smart Cities* **3**(3): 928–951.

[11] Yu, W., Xu, H., Nguyen, J., Blasch, E., Hematian, A. and Gao, W. (2018) Survey of public safety communications: User-side and network-side solutions and future directions. *Ieee Access* **6**: 70397–70425.

[12] Costin, A. (2013) Poor man's panopticon: Mass cctv surveillance for the masses. *PowerOfCommunity, November* .

[13] Mowery, K., Wustrow, E., Wypych, T., Singleton, C., Comfort, C., Rescorla, E., Halderman, J.A. *et al.* (2014) Security analysis of a full-body scanner. In *USENIX Security Symposium*: 369–384.

[14] Andrews, G. (2015) Police body cameras pre-installed with worm. `https://www.eteknix.com/police-body-cameras-pre-installed-worm/`.

[15] Muthusenthil, B. and Kim, H.S. (2018) Cctv surveillance system, attacks and design goals. *International Journal of Electrical and Computer Engineering (IJECE)* **8**(4).

[16] Olson, M. (2016) Beware, even things on amazon come with embedded malware. `http://artfulhacker.com/post/142519805054/beware-even-things-on-amazon-come`.

[17] Costin, A. (2016) Security of cctv and video surveillance systems: threats, vulnerabilities, attacks, and mitigations. In *Proceedings of the 6th International Workshop on Trustworthy Embedded Devices* (ACM): 45–54.

[18] Kharraz, A., Kirda, E., Robertson, W., Balzarotti, D. and Francillon, A. (2014) Optical delusions: A study of malicious qr codes in the wild. In *Dependable Systems and Networks (DSN), 2014 44th Annual IEEE/IFIP International Conference on* (IEEE): 192–203.

[19] Blasch, E.P., Rogers, S.K., Holloway, H., Tierno, J., Jones, E.K. and Hammoud, R.I. (2014) Quest for information fusion in multimedia reports. *International Journal of Monitoring and Surveillance Technologies Research (IJMSTR)* 2(3): 1–30.

[20] Grigoras, C. (2005) Digital audio recording analysis–the electric network frequency criterion. *International Journal of Speech Language and the Law* **12**(1): 63–76.

[21] Garg, R., Varna, A.L., Hajj-Ahmad, A. and Wu, M. (2013) "seeing" enf: power-signature-based timestamp for digital multimedia via optical sensing and signal processing. *IEEE Transactions on Information Forensics and Security* **8**(9): 1417–1432.

[22] Korycki, R. (2013) Time and spectral analysis methods with machine learning for the authentication of digital audio recordings. *Forensic science international* **230**(1-3): 117–126.

[23] Liu, Y., Yuan, Z., Markham, P.N., Conners, R.W. and Liu, Y. (2012) Application of power system frequency for digital audio authentication. *IEEE Transactions on Power Delivery* **27**(4): 1820–1828.

[24] Vidyamol, K. and George, J.P.J.E. (2017) Exploring electric network frequency for joint audio-visual synchronization and multimedia authentication. In *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)* (IEEE): 240–246.

[25] Hajj-Ahmad, A., Garg, R. and Wu, M. (2013) Spectrum combining for enf signal estimation. *IEEE Signal Processing Letters* **20**(9): 885–888.

[26] Rodríguez, D.P.N., Apolinário, J.A. and Biscainho, L.W.P. (2010) Audio authenticity: Detecting enf discontinuity with high precision phase analysis. *IEEE Transactions on Information Forensics and Security* **5**(3): 534–543.

[27] Ojowu, O., Karlsson, J., Li, J. and Liu, Y. (2012) Enf extraction from digital recordings using adaptive techniques and frequency tracking. *IEEE Transactions on Information Forensics and Security* **7**(4): 1330–1338.

[28] Chuang, W.H., Garg, R. and Wu, M. (2013) Anti-forensics and countermeasures of electrical network frequency analysis. *IEEE transactions on information forensics and security* **8**(12): 2073–2088.

[29] Vatansever, S., Dirik, A.E. and Memon, N. (2019) Factors affecting enf based time-of-recording estimation for video. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE): 2497–2501.

[30] Su, H., Hajj-Ahmad, A., Garg, R. and Wu, M. (2014) Exploiting rolling shutter for enf signal extraction from video. In *Image Processing (ICIP), 2014 IEEE International Conference on* (Citeseer): 5367–5371.

[31] Choi, J. and Wong, C.W. (2019) Enf signal extraction for rolling-shutter videos using periodic zero-padding. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE): 2667–2671.

[32] Vatansever, S., Dirik, A.E. and Memon, N. (2019) Analysis of rolling shutter effect on enf-based video forensics. *IEEE Transactions on Information Forensics and Security* **14**(9): 2262–2275.

[33] Vatansever, S., Dirik, A.E. and Memon, N. (2017) Detecting the presence of enf signal in digital videos: A superpixel-based approach. *IEEE Signal Processing Letters* **24**(10): 1463–1467.

[34] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P. and Süsstrunk, S. (2012) Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence* **34**(11): 2274–2282.

[35] Wang, S., Lu, H., Yang, F. and Yang, M.H. (2011) Superpixel tracking. In *2011 International Conference on Computer Vision* (IEEE): 1323–1330.

[36] Xu, Z., Min, B. and Cheung, R.C. (2019) A robust background initialization algorithm with superpixel motion detection. *Signal Processing: Image Communication* **71**: 1–12.

[37] Zivkovic, Z. and Van Der Heijden, F. (2006) Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern recognition letters* **27**(7): 773–780.

[38] Su, H., Hajj-Ahmad, A., Wong, C.W., Garg, R. and Wu, M. (2014) Enf signal induced by power grid: A new modality for video synchronization. In *Proceedings of the 2Nd ACM International Workshop on Immersive Media Experiences* (ACM): 13–18.