Modelling GOP Structure Effects on ENF-Based Video Forensics

Pasquale Ferrara(✉), Gerard Draper-Gil, Ignacio Sanchez, Henrik Junklewitz, and Laurent Beslay

European Commission – DG Joint Research Centre, Ispra, Italy
pasquale.ferrara@ec.europa.eu

Abstract. Electricity is transported through the network as alternate current, usually at a carrier frequency (50/60 Hz) which is known as Electric Network Frequency (ENF). In practice, ENF fluctuates around the nominal value because of changes in the supply and demand of power over time. These fluctuations are conveyed by the light that is emitted by sources connected to the power grid. Captured by video recordings, such localized variations can be exploited as digital watermarks in order to determine the position of a video in time (e.g. timestamping) and space, as well as to verify its integrity. However, the encoded formats of acquired videos alter the shape of ENF extracted from video frames. This paper provides an analytical model for characterizing the effects of group of pictures (GOP) structure adopted by the most widespread video encoders. The model is assessed through an experimental evaluation campaign, by analyzing different working conditions and by showing how the information from the GOP can contribute to the extraction of ENF from video frames.

Keywords: Electric network frequency · Video · Compression · GOP · Signal processing · Forensics

1 Introduction

As a result of the recent major digital transformation of the society, forensic analysis of multimedia evidence has acquired a growing importance in the investigation of crimes. Law enforcement investigation and prosecution have to face the growing misuse of new information technologies by criminals. Examples of such misuse include networks anonymization or data encryption, which defeat the efforts of the investigation process aiming to identify victims and perpetrators, to characterize the nature of the criminal activity and to localize the crime. In this context, multimedia forensics constitutes an increasingly central factor in criminal investigations. Audio and video forensic techniques [1] such as source camera identification or microphone forensics can enable law-enforcement to overcome the growing limitations posed of more classical forensic methodologies.

S. Goel et al. (Eds.): ICDF2C 2020, LNICST 351, pp. 121–138, 2021.
https://doi.org/10.1007/978-3-030-68734-2_7

Electric Network Frequency (ENF) analysis is another multimedia forensic technique that has demonstrated a strong potential to support forensic investigations[1]. When an electric device (computers, microphones, camcorders, surveillance cameras etc.) is plugged to the electric network, it captures not only a scene or a speech, but also the unique pattern of the 50/60 Hz electric network carrier frequency. ENF is not constant but it fluctuates around its nominal value because of the variations of power production and consumption. Such variations can be exploited as a digital watermark complementing the more classical forensic techniques used for timestamping and localization of multimedia recordings, and even overcome their limitations. These classical approaches rely on the analysis of content information which can be more and more subject to manipulation (spoofing, deep-fake, etc.). Metadata that might be present in the multimedia evidence (such as geo-tags or EXIF) could also contribute to the forensic analysis but they can be easily removed or tampered with by the perpetrators.

In 2010, ENF was considered as the "most significant development in audio forensics since Watergate"[2]. Indeed, ENF has been successfully used to extract forensic evidences from audio files that have been instrumental in court [2]. In particular, ENF forensics has proven to be effective with the following tasks:

- **Timestamping of the recording:** by matching the extracted ENF signal against a reference database of network frequencies of known electricity grids, it is possible to determine the precise point in time when the recording took place.
- **Localization of the recording:** it is possible to determine unique geographical constraints linked to the specific location, where the recording took place. This becomes possible by either analyzing specific features of the signal and comparing them statistically with intrinsic features present in each electric network or, by directly matching patterns extracted from recordings against a reference database of measurements of multiple registered electric networks from different geographical zones.
- **Integrity verification:** by analyzing the ENF signal extracted from the multimedia recording it is possible to identify fragments that have been edited, removed or inserted.

Most of the research ENF forensics has been focused on audio signals, where the ENF signal is included in the data stream due to the specific recording media used (e.g. audio recording in magnetic tapes) or electric induction over the power supply or other nearby strong magnetic fields.

More recently, researchers have demonstrated that the ENF signal can also be reconstructed from video recordings taken under artificial light, by exploiting the light fluctuations captured in the video [3]. The extraction of ENF from video frames is a new promising approach that, so far, has not yet been explored to its full potential. Among the factors that affects ENF-based video forensics [4], such as type of artificial light or time recording, there is still a lack of understanding on the role of video

[1] https://enfsi.eu/wp-content/uploads/2016/09/forensic_speech_and_audio_analysis_wg_-_best_practice_guidelines_for_enf_analysis_in_forensic_authentication_of_digital_evidence_0.pdf.

[2] https://www.theregister.co.uk/2010/06/01/enf_met_police/.

compression for the ENF captured in video frames. This aspect plays a key role as widely used smart devices provide only encoded videos. To the best of our knowledge, this work represents the first attempt to model the effects of group of pictures (GOP) structures, typical of the most common compression standard on the market such as H.262/MPEG-2, H.263, or newer ones such as H.264/MPEG-4 AVC and HEVC.

The rest of the paper is organized as follows: in Sect. 2 we provide an overview of the state-of-the-art on ENF-based media forensics. In Sect. 3, we focus on the extraction of ENF signal from video frames, while in Sect. 4 we explain the model that we propose to tackle with video compression. Then, we present our experimental results in Sect. 5, while we draw the conclusions in Sect. 6.

2 Related Works

The use of ENF in digital media forensics has gained large attention in the last decade due to its potential applications. Initial works focused on its application to digital audio forensics [5, 6], defining the "ENF criterion" as the reference procedure for the application of ENF in timestamping digital recordings. The ENF criterion procedure is based on three steps:

- ENF reference signal extraction (from the electricity grid);
- ENF audio signal extraction;
- ENF matching.

More recently, it has been demonstrated that ENF signals can also be extracted from video frames [3, 7] using the "light-flickering" effect, fluctuations of light intensity produced when illumination systems are connected to the power grid.

ENF reference signal extraction and matching are processing challenges independent of the source signal (i.e. audio or video frames). Most authors [8–10] assume that obtaining the ENF reference signal is a straightforward process, although the authors in [11] argued that ENF extraction from the electric grid is prone to errors if obtained from only one source.

Initially, ENF matching was done using visual inspection or Minimum Mean Square Error (MMSE) criteria [8, 12]. In the latest years, this process has captured the attention of several authors [13, 14], including proposals based on machine learning and more advanced statistical properties of ENF signals [15, 16].

There are many proposals for ENF extraction from audio signals [17–19], although according to [10], the improvements proposed in these works have a marginal effect on the overall result. Moreover, since most efforts address the problem of ENF extraction, other problems like ENF signal detection or audio tampering detection remain open.

Previous works on ENF extraction from video frames have followed two different approaches, those working with CCD [3, 7, 20] cameras and those working with CMOS [21–23] cameras. CCD sensor cameras adopt a global shutter system that capture all pixels at the same time, whereas CMOS sensor cameras use a rolling shutter sampling mechanism, capturing each pixel-row at different time instants. Moreover, other factors that can affect the quality of ENF signal are compression and light source.

The effects of rolling shutter have been studied in [3, 21, 23] with a comprehensive analysis done by the authors of [22]. The influence of the light source has also been addressed in [4]. Although, there are some papers addressing compression [3, 4], none of them offers a detailed explanation of its effects. Furthermore, none of the papers addressing compression envisage to take advantage of its effects and exploit them forensically.

3 ENF-Based Video Forensics

The voltage measured from the plug varies over time as:

$$V(t) = A_0 \cos(2\pi(f_0 + \varepsilon(t))t) \tag{1}$$

where f_0 is the nominal electrical network frequency (50 or 60 Hz, depending on the region) and $\varepsilon(t)$ models the fluctuation of the ENF over the time. Even though $\varepsilon(t)$ exhibits pseudo-periodic behavior due to the load control mechanism of the electricity grids [24], it can be considered as a unique pattern that allows localizing a given recording in time and space. In [7], authors demonstrated that ENF, defined as $f_0 + \varepsilon(t)$, impacts on most light sources by slightly changing the intensity of the emitted light. Being light intensity and electric voltage related to each other by a power law, light oscillations are at double frequency (100 or 120 Hz). It is also demonstrated that camera sensors can capture such variations during a video recording.

It is worth to note, however, that cameras usually acquire a video at a frame rate (25 or 30 frame per second) which is lower than light fluctuation (100 or 120 Hz). This implies that the ENF signal conveyed by the light is acquired at a sampling rate which is lower than the signal frequency, causing aliasing. This means that the ENF appears around a different frequency, which can be derived from the Nyquist theorem as:

$$f_a(t) = |f_x(t) - m \cdot f_s| \tag{2}$$

where $f_a(t)$, f_s and $f_x(t)$ are, respectively, the aliased frequency, the sampling frequency and the signal frequency; m is a positive integer such that $f_a(t) < f_s/2$. Assuming that $f_s = 30$ fps and $f_x(t) = 100 + \varepsilon(t)$ Hz, then $f_a(t) = |m \cdot 30 - 100 \cdot \varepsilon(t)| = 10$ Hz for $m = 3$.

Finally, modern cameras are equipped with camera sensors of different technologies, which employ different shutter technologies. CCD sensors are usually associated with a global shutter, meaning all pixels are illuminated at the same global time. CMOS sensors typically use a rolling shutter, so that not all pixels are acquired in the same time, but only each row. The latter could theoretically lead to a much higher sampling rate. However, in the rest of the paper we will consider only the global shutter approach, so that the ENF related signal $2f_0 + 2\varepsilon(t)$ is extracted by averaging all pixel intensities $I(i, j; t)$ for each video frame of size $M \times N$ as:

$$I(t) = \frac{1}{MN}\sum_{i,j=1}^{M,N} I(i,j;t). \tag{3}$$

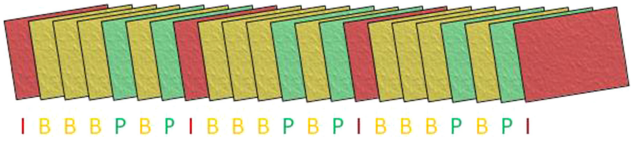

Fig. 1. Sequence of video frames whose GOP size is 7.

4 GOP Structure Effects on ENF Signal

We first recall some basics of video coding that are useful in order to comprehend the approach we propose for modelling the effects of compression when ENF is extracted from encoded videos.

4.1 Basics of Video Coding

By nature, every digital video is essentially a temporal sequence of pictures (i.e. frames) acquired at a given rate (e.g. 30 frames per second, in most of commercial devices). Such data stream can be properly compressed to save storage or transmission bandwidth by taking advantage of the redundancies that are present in the spatial (i.e. within each frame) and in the temporal (i.e. between two or more adjacent frames) domains and, at the same time, by exploiting the characteristics of the Human Vision System (HVS) to minimize visual degradation. To achieve this goal, the most common video compression standards such as MPEG-4 [25] and Advanced Video Coding (also known as MPEG-4 AVC or H.264) [26] employ a block-oriented and motion-compensated video coding approach. Briefly, such coding schemes divide frames into two different types: intra-coded frames, also referred as I-frames, and predictive-coded frames, which can in turn be divided in subtypes such as predicted (P) and bi-predicted (B) frames. During the encoding process, frames are grouped in GOPs (group of pictures) according to a structure that always begins with an I-frame and then present a certain number of predictive frames, as shown in Fig. 1. Encountering a new GOP in a stream means that the decoder does not need any previous frames to decode the next ones. The number of frames composing a group of pictures is called GOP size, and it might be constant or variable depending on the specific implementation.

When a frame is compressed, the encoder divides it in macroblocks (MBs) and encodes each MB individually: MBs belonging to I-frames are always encoded as they are, without referring to other frames, by means of a DCT quantization strategy. At the same time, MBs belonging to predictive-coded frames may be encoded referring to previous frames (this is the only possibility in P-frames) or even referring following frames (allowed in B-frames). Besides predicted MBs, the encoder embeds the motion vectors MVs associated to each MB. Finally, the encoder has also the possibility to skip a MB in a predictive-coded frame, if this MB can be directly copied from a previous

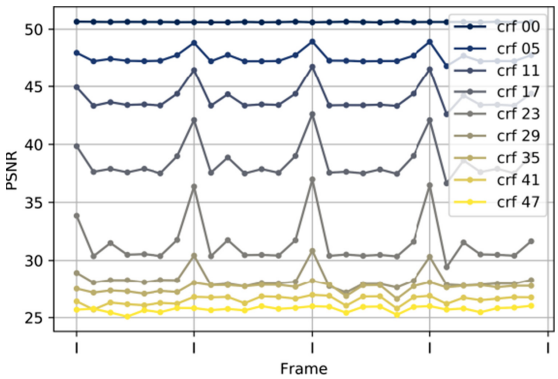

Fig. 2. PSNR between each raw frame and its corresponding encoded frame for different bitrates and a GOP size forced to 7.

frame (for instance, in presence of a static content). A more detailed description of video compression can be found in [27].

4.2 Modelling Compression Artifacts

We underline that, given their definitions in the last subsection, only I-frames certainly contain the information about pixels intensities of the original uncompressed frames, whereas in case of P and B-frames this information is borrowed from previous or neighboring frames.

In Fig. 2 we qualitatively illustrate the rationale of this assumption. We show the Peak Signal-to-Noise Ratio (PSNR) obtained between the uncompressed frames of a static video and their corresponding frames encoded for different constant rate factors (CRF). In this example, we forced the encoder to keep a GOP size equal to 7. For CRF $= 0$, the encoder codifies all frames as intra-coded at their maximum quality, so that to produce the average highest PSNR and no significant difference are visible among frames. Once the CRF increases, we observe that the fidelity of I-frames is systematically higher than the other predictive-encoded frames, which are more and mostly affected by the motion compensation strategy. For CRF greater than 29 (in our example), such differences between intra-coded and predictive-coded frames become smaller, because quantization effects on macro-block DCT coefficients become significant.

Starting from this observation, we model the way in which GOP structure acts on the video stream as a down-sampling operation. To simplify our analysis, we assume that a generic video is encoded by using a constant GOP size L, meaning that our intensity signal is subsampled of a factor L. From the Nyquist-Shannon theorem, we know that a band-limited signal can be reconstructed without *aliasing* if $F_S \geq 2B$, where F_S is the sampling frequency and B is the maximum frequency of the sampled signal. In our case, we can rewrite the Nyquist-Shannon inequality in function of the ENF frequency $F_X = 2f_0 + 2\varepsilon(t)$, the camera frame rate f_s and the GOP size L. First, we define the sampling rate after GOP encoding as:

$$F_S = \frac{1}{T_S^{GoP}} = \frac{1}{L \cdot \left(\frac{1}{f_s}\right)} = \frac{f_s}{L} \tag{4}$$

By substituting in the Nyquist-Shannon inequality, we obtain:

$$\frac{f_s}{L} \geq 2F_X \tag{5}$$

This means that aliasing due to the down-sampling operated by GOP structure does not appear if:

$$L \leq \frac{f_s}{2F_X} \tag{6}$$

As a numerical example, we consider the ENF extraction from a stand-still video using the global approach (Sect. 3) to obtain the signal $I(t)$. We assume that $f_0 = 50$ Hz and a camera frame rate $f_S = 30$ fps. Because the light flickering oscillates at a double frequency with respect the ENF, the aliased (see Sect. 3) ENF signal appears in the spectrum of the intensity signal at a frequency of $F_X = 10$ Hz. By applying these values to Eq. (6), we derive that the absence of the aliasing is guaranteed if and only if $L \leq 1.5$. On the contrary, for a GOP size $L \geq 2$, we observe aliasing in the spectrum of $I(t)$.

The assumption that the GOP size is constant can be relaxed by taking advantages of the generalization of the Nyquist-Shannon to nonuniform sampling [28], by considering the average sampling period, or GOP size in this specific application.

Finally, the Nyquist-Shannon theory also provides information about the positions of the alias ENF in the Fourier spectrum. By using the symmetry property of the Fourier transform of real signals, the location of aliased ENF signals is:

$$F_X^a = \begin{cases} F_X - k\frac{f_S}{L} & \text{if } 0 < F_X - k\frac{f_S}{L} < f_s/2 \\ \mathrm{mod}\left(f_s/2 - \left(F_X - k\frac{f_S}{L}\right), f_s/2\right) & \text{elsewhere} \end{cases}, \; k \in Z. \tag{7}$$

It is worth to note that for $F_X - k\frac{f_S}{L} \notin (0, f_s/2)$, the alias spectrum is $I(-f) = -I(f)$. In the case of variable GOP size, the model still applies but with $\bar{L} = \frac{1}{N}\sum_{i=0}^{N-1} L_i$, where L_i are the GOP sizes of each group of pictures present in video.

5 Results and Discussions

This Section provides an experimental analysis validating the model we described in Sect. 4, in order to demonstrate its capability to predict the effects of video coding when extracting ENF signal from video frames.

5.1 Experimental Settings

We employed an Allied Vision[3] Manta GigaE 145BNIR camera to acquire uncompressed video frames of a white wall scene. The camera is equipped with a CCD sensor whose resolution is 347 × 259 pixels and adopting a global shutter system. We set the frame rate at 30 frame per second, constant shutter time and manual focus. Each frame is transferred to and stored into a laptop by means of a Gigabit Ethernet connection, in a bitmap format.

Two uncompressed videos of at least 3 min are acquired by using sunlight and indoor LED light powered by the electrical network. The nominal electric network frequency is 50 Hz. While acquiring the video with artificial light, we also recorded a reference ENF directly from the electric plug, in order to check whether any recorded signal extracted in the video is actually the real ENF or not.

From these two raw videos, we generated compressed videos by making use of FFMPEG[4], a well-known tool for video processing and coding. In terms of type compression, we divided our experiments into three stages: first, we analyze the case constant GOP size; then, we carried out our analysis for variable GOP size and, finally, we analyze the case of variable bitrates. In all cases, we applied different GOP sizes and/or different bitrates.

The analysis is carried on in the Fourier domain, by analyzing the periodograms and the spectrograms of the signal obtained as Eq. (3). For the periodogram based analysis, we used 512 FFT points and Hamming window. For the spectrogram-based analysis, we used 8192 FFT points, Hamming window, 30 s of frame time and 29/30 of overlap between frames in order to have a time resolution of 1 s.

5.2 Reference ENF Acquisition

We used an Arduino UNO with an emonTxShield[5] to sample the voltage from the electric grid at a sampling rate of 1 kHz. The raw voltage values are directly measured from the emonTxShield. In order to extract the ENF, the voltage measurements can be analyzed in the frequency (periodogram) and time-frequency (spectrogram) domains by means of Short-time Fourier transform (STFT).

The resulting reference ENF signal is shown Fig. 3. From the periodogram figure (first subplot), we observe that most of the energy is concentrated around the fundamental frequency f_0 = 50 Hz, while other peaks are visible at its integer multiples (harmonics). By looking at the spectrogram (second subplot in Fig. 3), centered at the fundamental frequency, we observe that the signal energy (yellow) is deviating from the nominal value over time. Finally, the network frequency at a given time can be estimated from the spectrogram by detecting the energy peak along frequencies axis, for instance by using quadratic interpolation [12]. The punctual ENF is show in the third subplot of Fig. 3.

[3] https://www.alliedvision.com/en/digital-industrial-camera-solutions.html.

[4] https://ffmpeg.org/.

[5] https://wiki.openenergymonitor.org/index.php/EmonTx_Arduino_Shield.

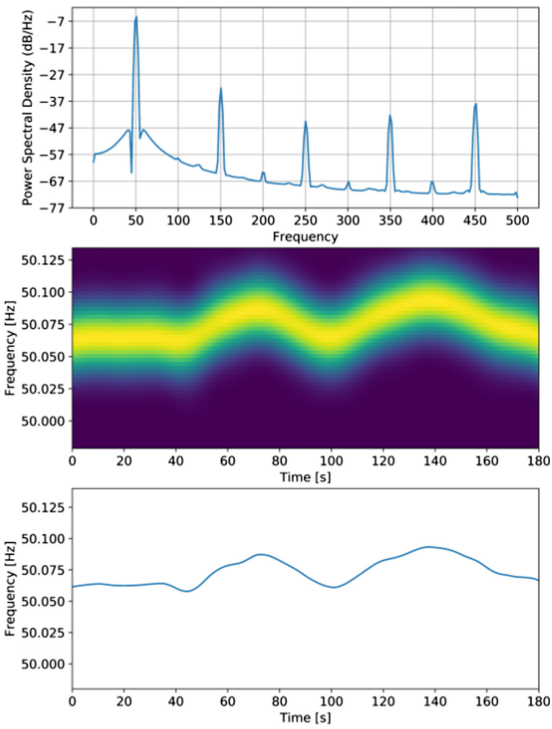

Fig. 3. Power spectral density, periodogram and punctual values of the reference ENF measured by an electric plug. (Color figure online)

5.3 ENF Extraction from Raw Frames

For each frame, the average pixels intensity is calculated as in Eq. (3), and the resulting 1-D signal is processed to extract ENF. As we have done for the reference ENF signal, we show the periodograms and the spectrograms of I(t) for the two uncompressed recordings that we have acquired with sunlight and artificial light illumination. The results are presented is Fig. 4. In the case of artificial light (a), we clearly see a prominent peak at 10 Hz in the periodogram because of the aliasing that raises with these experimental settings. Moreover, we can visually recognize the same ENF time-frequency behavior in the spectrogram between the refence ENF (Fig. 3) and the one extracted from the video (Fig. 4(a)). Conversely, for the signal extracted under sunlight condition (Fig. 4(b)), we do not observe any peak either in the periodogram (apart from the DC component) or in the spectrogram.

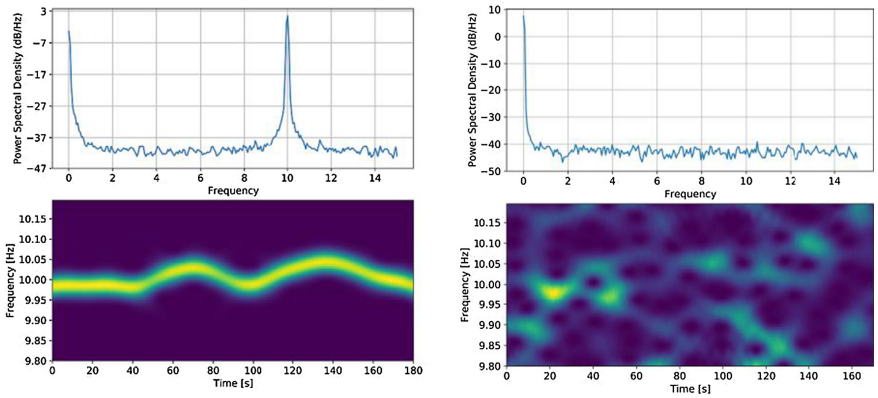

Fig. 4. Periodograms and spectrograms of the average frame intensities from uncompressed videos under LED indoor (a) and sunlight (b) illumination. In (a) ENF is present at the expected frequency, whereas in (b) it is not present.

5.4 Constant GOP Size

First, we compressed both raw recordings under artificial light and sunlight illumination with different constant GOP size L. We did it through FFMPEG by using H.264 coding, a chroma factor yuv420p and by forcing the encoder to adopt a fixed GOP size. We considered 8 GOP sizes L = {1, 7, 11, 15, 30, 60, 135, 222}. Then, we analyzed the signal I(t) for each video. For convenience, we split our analysis between $L < f_s$ and $L \geq f_s$.

In Fig. 5 we globally note that, despite different GOP sizes, the peak at the fundamental frequency 10 Hz is maintained for all GOP sizes, and that its SNR does not vary significantly. However, depending on L, we assist to the emergence of other small peaks at different positions in function of L, except for the case of L = 1. Such peaks can be explained by the aliasing effects due to the GOP structure. We confirm this

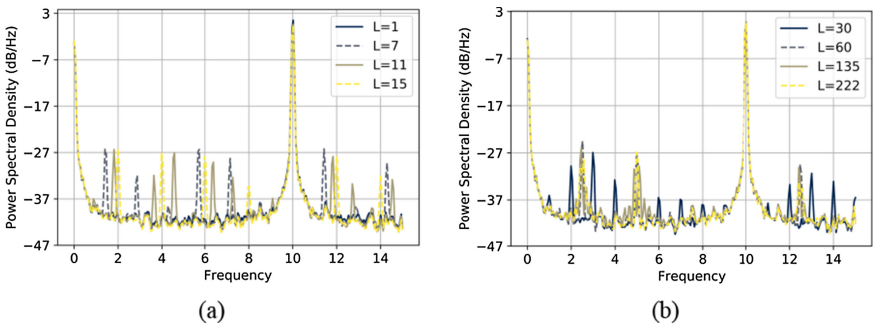

Fig. 5. Periodograms of I(t) extracted from videos encoded with constant GOP sizes, either if $L < f_s$ (a) or $L \geq f_s$ (b), with LED artificial light.

deduction in Fig. 6, wherein we plot the spectrograms of the signal for three different GOP sizes (11, 30, 135). To better visualize them, we cropped the complete spectrogram in correspondence of some expected alias frequencies. For instance, for L = 11 we should observe the aliased signal at $10-\mathrm{k}\left(\frac{30}{11}\right) = 10-\mathrm{k}\cdot 2.73$. This hypothesis is confirmed in Fig. 6(a), where we observe distinctive replicas of the same signal at

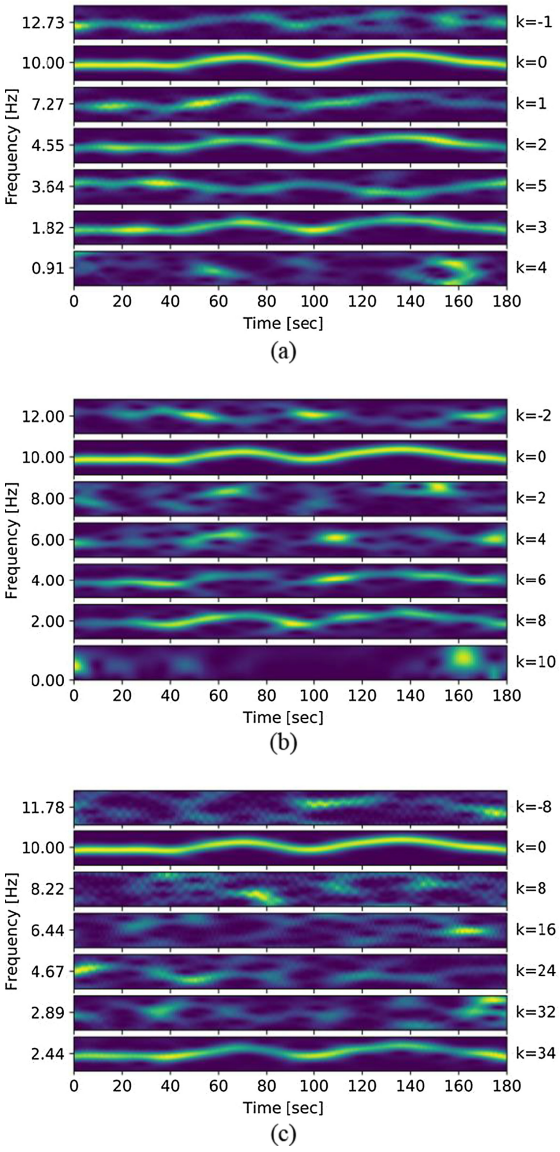

Fig. 6. Spectrograms for GOP size L = 11 (a), L = 30 (b) and L = 135 (c). For each subfigure, we cropped the fundamental component (k = 0) and the most significant aliased frequency.

4.55 Hz (k = 2) and 1.82 Hz (k = 4). Similar considerations can be drawn from the other spectrograms, where we observe temporally correlated energies at 2 Hz for L = 30 and 2.44 Hz for L = 135. It is also worth to note that the alias components do not have the same SNR. Moreover, some of them are in opposite phase with respect to the principal component (see the case k = 3 for L = 11).

We repeated the same analysis for the video that does not contain the ENF signal, since it is acquired with sunlight. The periodograms are shown in Fig. 7, and they do not present any prominent peak, as expected, apart from the DC component. However, we observe that for L > 1 there are several spurious peaks whose position changes with L. This phenomenon is still in agreement with our model, but with respect to the previous case, what is aliased now is the DC components. In fact, we can verify that every peak is spaced of d = 30/L starting from frequency 0.

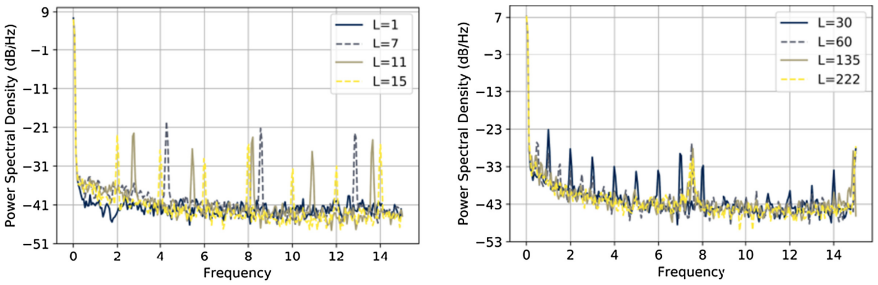

Fig. 7. Periodograms of average frame intensity signal extracted from videos encoded with constant GOP sizes, either if $L < f_s$ (a) or $L \geq f_s$ (b), with sunlight.

This consideration is confirmed in Fig. 8, where we show the spectrogram of the signal for L = 7 by cropping it at the expected alias frequency. Thus, we verify that effectively the prediction of the alias position is correct and that the signal is almost constant over the time being related to DC components.

5.5 Variable GOP Size

In this experimental setting, we address the case in which the video stream is encoded with a variable GOP size. To do that, we built up an experiment that consists of producing a set of videos at a fixed constant bitrate (CRF = 23 is the standard quality in FFMPEG7H.264) but forcing the encoder to adopt a GOP size variable but close to predetermined values. The parameters are summarized in Table 1.

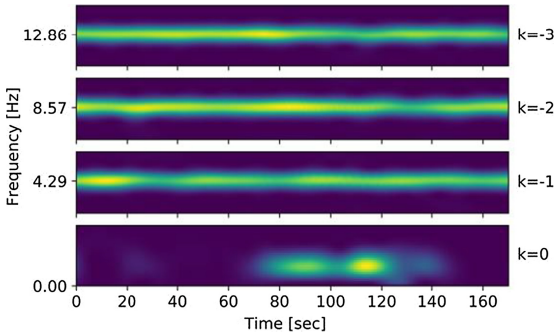

Fig. 8. Spectrogram obtained for L = 7 from the video acquired with sunlight.

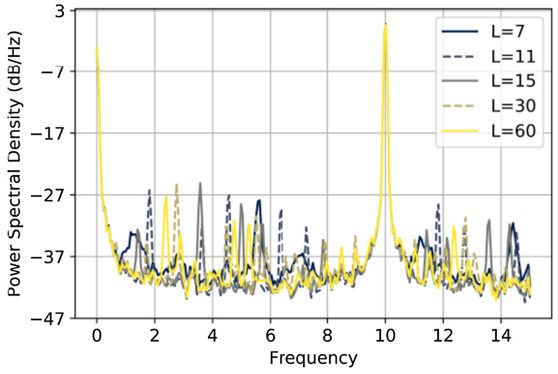

Fig. 9. Periodograms of I(t) extracted from videos encoded with variable GOP size.

We extracted the actual GOP structure from the video streams by means of the FFMPEG routine *ffprobe*. As in the previous set of experiments, we analyzed the effects of GOP structure by means of periodograms and spectrograms. In Fig. 9, we plotted the power density spectra of the intensity signals I(t), for each case study. In a similar way to the case of constant GOP size, several alias frequencies emerge in the periodogram. However, we registered the main differences in the spectrograms. In Fig. 10, we show the spectrograms for L close to 11, in subfigure (a), and for L close to 30, in subfigure (b), analogously to Fig. 6(a) and Fig. 6(b). By comparing Fig. 10(a) and Fig. 6(a), we observe that patterns are remarkably similar each other. At the same time, a similar conclusion cannot be drawn by observing Fig. 10(b) and Fig. 6(b). We can therefore conclude that the model has less predictive power in the case of variable GOP size, even though for some configurations it still applies satisfactorily.

Table 1. Parameters adopted in variable GOP size experiment. Nominal and real (average) GOP size are highlighted, as well as the bitrates of the resulting videos.

CRF	Nominal L	Average L	Bitrate
23	7	6.796	2383 kbps
23	11	10.995	2460 kbps
23	15	13.974	2408 kbps
23	30	28.759	2322 kbps
23	60	57.843	2246 kbps

5.6 Variable Bitrates with Different GOP Sizes

In this last test we analyze the case of different bit rates given a fixed GOP size. We forced the encoder to adopt two different GOP sizes, 7 and 30 and we varied the constant rate factor CRF = {0, 5, 11, 17, 23, 29, 35, 41, 47}. The periodograms of the signals extracted from each video are shown in Fig. 11. As we expected, aliased patterns appear also in this case, but we assist also to a lessening of the ENF SNR at 10 Hz when the bitrate decreases. The most interesting part of this analysis is provided in Fig. 12, where we show spectrograms cropped at the frequencies where the aliased peaks have the higher power spectral densities. For the sake of simplicity, we show them for CRF = 23 (standard compression in FFMPEG) and CRF = 47 (heavy compression). It can be noticed that for L = 7, the ENF signal energy is concentrated at 10 Hz and non-significant replicas are present, independently from the bitrate. Conversely, for L = 30, weak aliased signals are visible, especially at 11.04 Hz and the replicas become even more pronounced when the bitrate decrease (see bottom-right plot of Fig. 12). This behavior can be explained qualitatively as follows: at a given bitrate, we forced the encoder to have a certain number of I-frames, accordingly to the GOP size; because of the constraint on the total bitrate, the DCT coefficients quantization applied to the I-frames is strong for small GOP sizes (i.e. high number of I-frame for a given recording). This fact negatively affects the aliased frequencies making them not distinguishable, even at high bitrates. On the other hand, when the GOP size is larger, so that the number of I-frames encoded decreases, the DCT coefficients quantization becomes less severe. This means that I-frames have a higher PSNR compared to the case of small GOP size, so that the effects of GOP structure adhere more to the down-sampling model that we described in Subsect. 4.2.

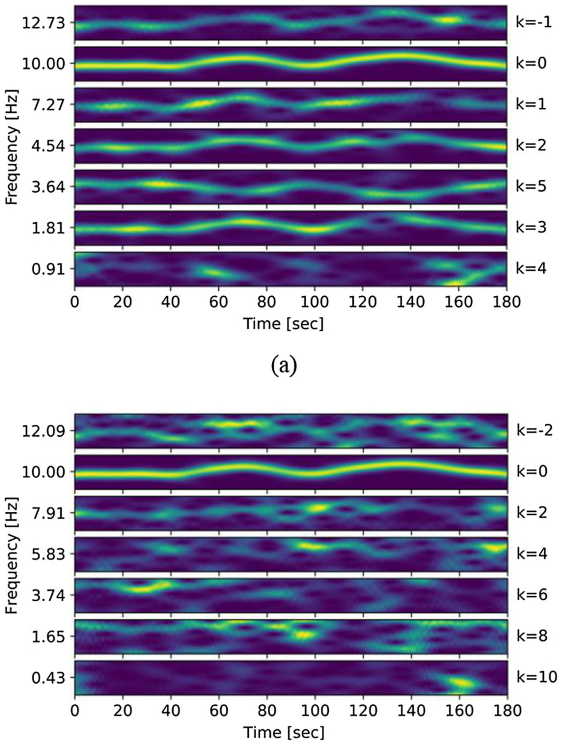

Fig. 10. Spectrograms for variable GOP size L = 11 (a), L = 30 (b). For each subfigure, we cropped the fundamental component (k = 0) and some aliased frequency for $k \neq 0$.

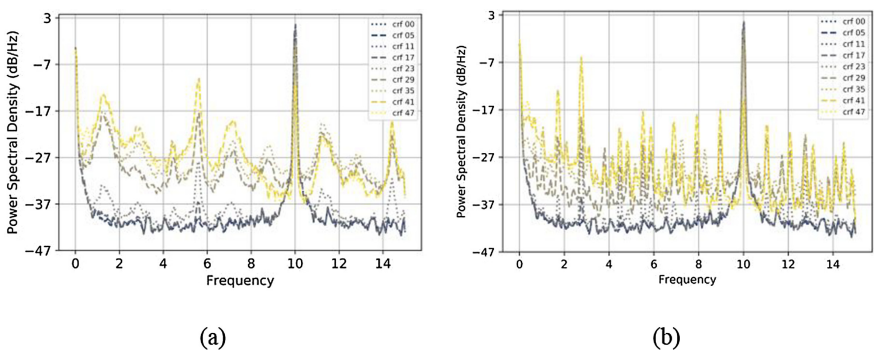

Fig. 11. Periodograms of I(t) extracted from videos encoded with L = 7 (a) and L = 30 (b) at different constant rate factors.

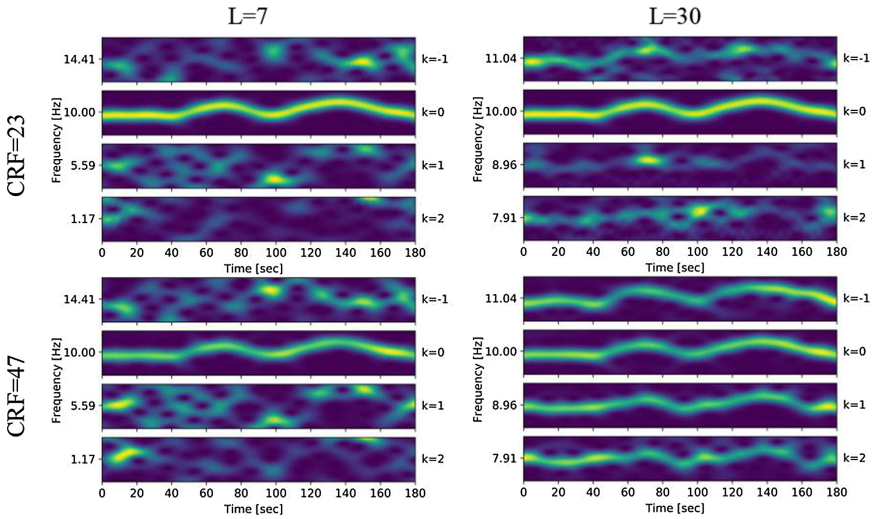

Fig. 12. Spectrograms of intensities signals for 4 combination of GOP sizes (7 and 30) and different CRF (23 and 47).

5.7 Discussions and Applications

The main findings of our analysis can be summarized as follows:

1. At a given bitrate and constant GOP size, the fundamental ENF is present and its position and SNR do not depend on the GOP size.
2. Alias ENF replicas appear in the spectrum in function of the GOP size with a lower SNR, matching out theoretical predictions within reasonable bounds.
3. The SNR of aliased ENFs decreases if the GOP size is variable and, generally, by increasing the GOP size.
4. The SNR of the fundamental ENF decrease with the video bit rate.

In a forensic setting, the analyst knows the video frame rate and its GOP structure, if the video is not re-encoded or tampered. We can also assume that the nominal frequency is known (only two cases are possible, 50 or 60 Hz). From these postulates, we can propose several approaches in order to take advantages of this analysis. As we have seen in our experiments, some GOP configurations make ENF emerge in other bands so that, by taking into consideration these effects, ENF signal detection or estimation can be easier and more reliable. For instance, the ENF signal can be extracted more reliably by combining those alias frequencies with a good SNR [18]. Or they can be used to target Multiple Signal Classification (MUSIC) based techniques [29], wherein the signal power spectrum is estimated by defining in advance the number of sinusoidal components (i.e. eigenspace signal decomposition). Therefore, specifying the correct number of sinusoidal components, that might be present within the signal because of compression effects, could lead to better spectral estimation. Finally, it can also help when the nominal ENF is a multiple integer of the frame rate and, therefore, ENF appears around the DC component.

6 Conclusions

In this paper, we modelled the effects of GOP structure on the ENF signal conveyed by the light and then captured by a video camera. We formalized the problem by making use of Nyquist-Shannon sampling theory and we showed how side information such as video frame rate, targeted ENF and GOP size allows to predict the spectrum of the signal containing ENF. As a new and determinant contribution for improving forensic procedure, our study demonstrated that the ENF signal is affected by the aliasing due to the GOP structure, with different degrees in function of the GOP size and if this last is constant or variable, as well as in function of the total bitrate. In future works, moving closer to videos processed in real investigations, we will take advantage of the outcomes presented in this work to design stronger ENF detection and extraction procedures, and apply this methodology to video time-stamping and geo-localization.

References

1. Ho, A.T.S., Li, S.: Handbook of Digital Forensics of Multimedia Data and Devices, 1st edn. Wiley – IEEE Press, Hoboken (2015)
2. Kajstura, M., Trawinska, A., Hebenstreit, J.: Application of the electrical network frequency (ENF) criterion: a case of a digital recording. Forensic Sci. Int. **155**(2), 165–171 (2005)
3. 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 Trans. Inf. Forensics Secur. **8**(9), 1417–1432 (2013)
4. Vatansever, S., Dirik, A.E., Memon, N.: Factors affecting ENF Based time-of-recording estimation for video. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 2497–2501. IEEE, Brighton (2019)
5. Grigoras, C.: Digital audio recording analysis: the electric network frequency (ENF) criterion. Int. J. Speech Lang. Law **12**(1), 63–76 (2005)
6. Grigoras, C.: Applications of ENF criterion in forensic audio, video, computer and telecommunication analysis. Forensic Sci. Int. **167**(2), 136–145 (2007)
7. Garg, R., Varna, A.L., Wu, M.: 'Seeing' ENF: natural time stamp for digital video via optical sensing and signal processing. In: Proceedings of the 19th ACM International Conference on Multimedia, pp. 23–32. Association for Computing Machinery, New York (2011)
8. Ojowu, O., Karlsson, J., Li, J., Liu, Y.: ENF extraction from digital recordings using adaptive techniques and frequency tracking. IEEE Trans. Inf. Forensics Secur. **7**(4), 1330–1338 (2012)
9. Hua, G., Goh, J., Thing, V.L.L.: A dynamic matching algorithm for audio timestamp identification using the ENF criterion. IEEE Trans. Inf. Forensics Secur. **9**(7), 1045–1055 (2014)
10. Hua, G., Bi, G., Thing, V.L.L.: On practical issues of ENF based audio forensics. IEEE Access **3536**, 20640–20651 (2017)
11. Elmesalawy, M.M., Eissa, M.M.: New forensic ENF reference database for media recording authentication based on harmony search technique using GIS and wide area frequency measurements. IEEE Trans. Inf. Forensics Secur. **9**(4), 633–644 (2014)
12. Cooper, A.: An automated approach to the electric network frequency (ENF) criterion: theory and practice. Int. J. Speech Lang. Law **16**(2), 193–218 (2009)

13. Hua, G.: Error analysis of forensic ENF matching. In: IEEE International Workshop on Information Forensics and Security, pp. 1–7. IEEE, Hong Jong (2018)
14. Zheng, L., Zhang, Y., Lee, C.E., Thing, V.L.L.: Time-of-recording estimation for audio recordings. Digit. Invest. **22**, S115–S126 (2017)
15. Hajj-Ahmad, A., Garg, R., Wu, M.: ENF-based region-of-recording identification for media signals. IEEE Trans. Inf. Forensics Secur. **10**(6), 1125–1136 (2015)
16. Cui, Y., Liu, Y., Fuhr, P., Morales-Rodriguez, M.: Exploiting spatial signatures of power ENF signal for measurement source authentication. In: IEEE International Symposium on Technologies for Homeland Security, pp. 1–6. IEEE, Woburn (2018)
17. Bykhovsky, D., Cohen, A.: Electrical network frequency (ENF) maximum-likelihood estimation via a multitone harmonic model. IEEE Trans. Inf. Forensics Secur. **8**(5), 744–753 (2013)
18. Hajj-Ahmad, A., Garg, R., Wu, M.: Spectrum combining for ENF signal estimation. IEEE Signal Process. Lett. **20**(9), 885–888 (2013)
19. Su, H, Hajj-Ahmad, A., Wong, C.-W., Garg, R., Wu, M.: ENF signal induced by power grid: a new modality for video synchronization. In Proceedings of the 2nd ACM International Workshop on Immersive Media Experiences. Association for Computing Machinery, New York, pp. 13–18 (2014)
20. Vatansever, S., Dirik, A.E., Memon, N.: Detecting the presence of ENF signal in digital videos: a superpixel-based approach. IEEE Signal Process. Lett. **24**(10), 1463–1467 (2017)
21. Vatansever, S., Dirik, A.E., Memon, N.: Analysis of rolling shutter effect on ENF-based video forensics. IEEE Trans. Inf. Forensics Secur. **14**(9), 2262–2275 (2019)
22. Su, H., Hajj-Ahmad, A., Garg, R., Wu, M.: Exploiting rolling shutter for ENF signal extraction from video. In 2014 IEEE International Conference on Image Processing, pp. 5367–5371. IEEE (2014)
23. Hajj-Ahmad, A., Berkovich, A., Wu, M.: Exploiting power signatures for camera forensics. IEEE Signal Process. Lett. **23**(5), 713–717 (2016)
24. Lin, X., Kang, X.: Supervised audio tampering detection using an autoregressive model. In IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2142–2146. IEEE, New Orleans (2017)
25. ISO: Information technology - coding of audio-visual objects-part 2: Visual. ISO/IEC IS 14496-2, International Organization for Standardization, Geneva (2009)
26. ISO: Information technology - coding of audio-visual objects - part 10: Advanced video coding (AVC). ISO/IEC IS 14496-10, International Organization for Standardization, Geneva (2010)
27. Bovik, A.C.: Handbook of Image and Video Processing, 2nd edn. Elsevier, Amsterdam (2005)
28. Marvasti, F.: Nonuniform Sampling, Theory and Practice. Springer, Kluwer Academic/Plenum Publishers, New York (2001)
29. Schmidt, R.: Multiple emitter location and signal parameter estimation. IEEE Trans. Antennas Propag. **34**(3), 276–280 (1986)