EXPLOITING ROLLING SHUTTER FOR ENF SIGNAL EXTRACTION FROM VIDEO

Hui Su, Adi Hajj-Ahmad, Ravi Garg, and Min Wu
University of Maryland, College Park
{hsu, adiha, ravig, minwu}@umd.edu

ABSTRACT

The electric network frequency (ENF) signal can be embedded in multimedia recordings created in areas of electrical activities. Recent work has used the ENF signal for such applications as time stamp authentication and forgery detection. It is more challenging to extract ENF signals from video recordings than from audio recordings because of the low temporal sampling rate or frame rate of video cameras. The rolling shutter of CMOS image sensor can be exploited as it exposes a frame line by line, and the effective ENF sampling rate by treating each line as a signal sample can be increased. This scheme was shown to work well with static videos. This paper conducts a further study on the exploitation of the rolling shutter for extracting ENF traces from videos. The rolling shutter mechanism is modeled and analyzed using multirate signal processing theory. Challenging cases of videos with motions are examined, and solutions to extracting ENF from them are explored.

Index Terms— ENF, rolling shutter, video, filter bank

1. INTRODUCTION

Electric Network Frequency (ENF) based analysis has emerged in recent years as a promising tool for such multimedia forensic applications as time stamp authentication and forgery detection. ENF is the frequency of the alternating current in an electric power grid. The nominal value of the ENF is 60Hz (in North America) or 50Hz (in most other parts of the world). The actual value of the ENF fluctuates slightly around its nominal value over time as a result of the interaction between power load and generation, and the main trends of these fluctuations are consistent within the same power grid. The changing values of the ENF over time are regarded as an ENF signal. One can extract the ENF signal from measurements at a power outlet using a step-down transformer and a voltage divider circuit.

Interestingly, audio recordings created using devices plugged into the power mains or located near power sources can pick up the ENF signals due to electromagnetic interference or acoustic vibrations [1]. Several forensic applications have been proposed based on the analysis of the ENF traces in audio recordings. In [1, 2], the ENF signal is used as a natural time stamp to authenticate audio recordings. In [3], the authors propose to detect the region of tampering by examining the phase continuity of the ENF signal. It is shown in [4, 5, 6] that the ENF signal can also reveal information about the locations and regions in which certain recordings are made. Applications of ENF analysis beyond forensics have also been proposed [7].

Most previous work related to the analysis of ENF signals is built on extracting ENF traces from audio recordings [1, 2, 3, 8]. Recently, it has been found that indoor lightings such as the fluorescent lights and incandescent bulbs vary their light intensity in accordance with the AC voltage supplied, which varies according to the AC supply frequency [9]. As a result, it is possible for cameras to capture the light intensity variation that can be used to extract the ENF signal. In [9], the authors took the mean of the pixel values in every frame of video recordings that capture indoor lightings, and then used spectrogram analysis to estimate the embedded ENF signal. One major challenge of this scheme is the aliasing effect. Most of the consumer digital cameras adopt a frame rate of around 30 fps, while the ENF signal appears at harmonics of 50 or 60 Hz. Therefore the ENF signal suffers from severe aliasing effect induced by insufficient sampling speed. In the US, the nominal value of the ENF is 60 Hz. If the frame rate is exactly 30 Hz, the ENF signal will be shifted to 0 Hz, i.e., the DC frequency. As a result, it is very difficult to estimate the ENF signal due to low signal-to-noise ratio. Figure 1 shows the spectrogram calculated from the mean values of the frames of a video shooting a white wall under fluorescent lightings. We can see the ENF signal overlaps with the DC components and is difficult to extract.

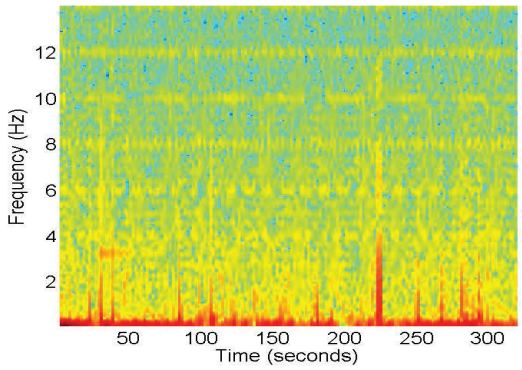


Fig. 1. The spectrogram of a video recording shooting a white wall under under fluorescent lightings. Figure is best viewed in color.

In our previous work, we have proposed to take advantage of the rolling shutter to address the problem of insufficient sampling rate [10]. The rolling shutters are commonly adopted for the complementary metal-oxide semiconductor (CMOS) camera sensors. Unlike cameras with global shutters that record an entire frame from a snapshot of a single point in time, a camera with a rolling shutter scans the vertical or horizontal lines of each frame in a sequential manner, so that different lines in the same frame are exposed at different times. If we treat each line of the frame as a sample, the temporal sampling rate can be much higher than the frame rate, which would facilitate the estimation of the ENF signal.

In this work, we carry out a further study on the exploitation of the rolling shutter for extracting ENF traces from video recordings. We model and analyze the rolling shutter mechanism with a

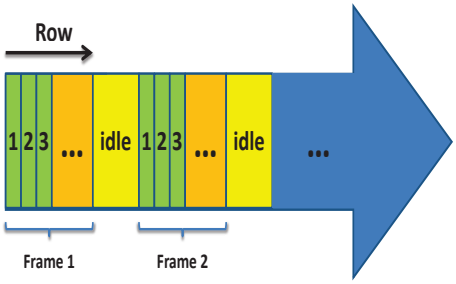
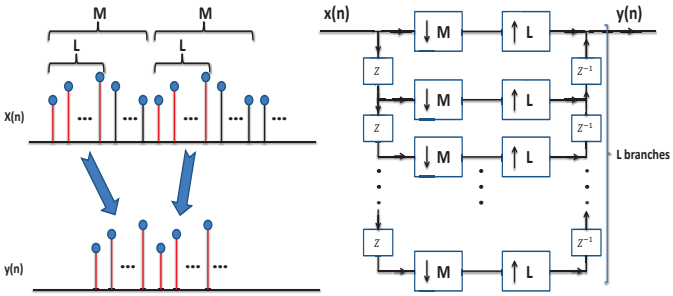


Fig. 2. Timing of rolling shutter sampling: the rows of a frame are sequentially exposed, followed by an idle period before proceeding to the next frame.



(a) (b)

Fig. 3. (a) Time domain illustration of the rolling shutter sample acquisition; (b) an equivalent filter bank model ($L \leq M$).

filter bank using multirate signal processing theory. We then extend the scope of extracting ENF traces from videos of still scenes to those containing motions, which is a challenging problem and has never been formally attempted. Several methods are developed and promising results are observed.

2. THE ROLLING SHUTTER

With a rolling shutter, each frame is recorded by scanning across the frame either vertically or horizontally line by line, instead of capturing the whole frame at a single point in time as in the case of a global shutter. Figure 2 illustrates the timing for the image acquisition of rolling shutters, assuming the scanning of a frame is done row-by-row. Each row of the frame is sequentially exposed to light, followed by a possible idle period before proceeding to the next frame. Since pixels in different rows are exposed at different times but are displayed simultaneously during playback, the rolling shutter may cause such distortions as skew, smear, and other image artifacts, especially with fast-moving objects and rapid flashes of light [11].

The sequential read-out mechanism of rolling shutter has been traditionally considered detrimental to image/video quality due to its accompanying artifacts. However, recent work has shown that the rolling shutter can be exploited with computer vision and computational photography techniques [12, 13]. Our previous work [10] has exploited the rolling shutter to extract ENF traces from videos of static scenes. In this paper, we investigate the more challenging cases of videos containing motions.

A Filter Bank Model

For an image captured by a rolling shutter, we can treat the spatial mean of every row as a temporal sample since all the pixels in a single row are exposed at the same time instance. As there is a between-frame idle period during which no rows are exposed, in terms of capturing the ENF signal over time we are equivalently abandoning some samples that would have been generated in the idle period. The time domain illustration of this model is shown in Figure 3 (a). Here, we assume that the shutter is able to produce M samples at its full capacity, and only L samples among them are retained while the rest are discarded, where $L \leq M$. We denote the input and output signal as $x(n)$ and $y(n)$, respectively.

To facilitate frequency domain analysis, we use a L-branch filter bank [14] to model the relationship between the input signal $x(n)$ and the output signal $y(n)$, as shown in Figure 3 (b). In each branch of the filter bank, the input goes through an M-fold down-sampler followed by an L-fold up-sampler, with appropriate delays at both the beginning and the end of the branch.

The DTFT of the signal coming out of the l^{th} branch can be analyzed according to multi-rate signal processing theory [14]:

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

So the DTFI of the combined final output $Y(\omega)$ is given by:

$$\begin{aligned} Y(\omega) &= \sum_{l=0}^{L-1} Y_l(\omega) \\ &= \sum_{l=0}^{L-1} \frac{1}{M}\left(\sum_{m=0}^{M-1} X(\frac{\omega L + 2\pi m}{M}) e^{j\frac{\omega L+2\pi m}{M}l}\right) e^{-j\omega l} \\ &= \sum_{m=0}^{M-1} X(\frac{\omega L + 2\pi m}{M}) F_m(\omega), \end{aligned} \quad (2)$$

where

$$F_m(\omega) = \frac{1}{M}\sum_{l=0}^{L-1} e^{-j\frac{\omega(M-L)-2\pi m}{M}l}. \quad (3)$$

3. EXTRACTION OF ENF TRACES

In this section, we describe how to extract ENF traces from videos captured by rolling shutters. Given a video signal, we calculate the spatial mean of the pixel values for every row, and refer to it as a row signal hereafter. The row signal contains mainly two components: the video signal corresponding to the visual scene and the ENF signal. Denote by $R(r, n)$ the row signal from the r^{th} row in the n^{th} frame, we have

$$R(r, n) = V(r, n) + E(r, n), \quad (4)$$

where $V(r, n)$ is the visual signal, and $E(r, n)$ is the ENF signal.

Denote by T the frame duration of the camera. According to the notations in the previous section, the sampling rate of the shutter is $f_s = M/T$, and the perceptual sampling rate of the row signal is L/T. Here, L is the number of rows per frame, and the exact value of M depends on the CMOS manufacture's design and is usually unknown to the public.

The spectrogram of the row signal is computed using the perceptual sampling rate L/T, instead of the actual sampling rate M/T.

Deriving from Equation (2), we can show that the Fourier representation of the row signal is

$$Y(f) = \sum_{m=0}^{M-1} X(\frac{2\pi}{f_s}(f + \frac{m}{T}))F_m(f). \quad (5)$$

This suggests that the row signal is the weighted summation of a series of transformed versions of $x(n)$ that are shifted by multiple of $\frac{1}{T}$ in the frequency domain.

3.1. Static Video

We start with the simplest case in which the scene in the video recording is constant, so that the video signals of every frame in the video are identical. Equation (4) is thus reduced to

$$R(r, n) = V(r) + E(r, n). \quad (6)$$

Note here $E(r, n)$ is sampled from a sinusoid signal whose frequency is deviating slightly from its nominal value. So the average of $E(r, n)$ for a given row r across a large number of frames (e.g., 100) should be close to 0, i.e.

$$\bar{E}(r) = \frac{\sum_n E(r, n)}{\sum_n 1} \simeq 0. \quad (7)$$

Subtracting from $R(r, n)$ its average value across the frames, denoted as $\bar{R}(r)$, we have

$$\begin{aligned} \hat{R}(r, n) &= R(r, n) - \bar{R}(r) \\ &= R(r, n) - \frac{\sum_n R(r, n)}{\sum_n 1} \\ &\simeq R(r, n) - V(r) = E(r, n). \end{aligned} \quad (8)$$

We can then estimate the ENF signal from $\hat{R}(r, n)$. The ENF signal-to-noise-ratio (SNR) of $\hat{R}(r, n)$ is higher than that of $R(r, n)$, leading to a more robust and accurate estimation.

We conducted several experiments using a Canon PowerShot SX230 camera that has a CMOS sensor with a rolling shutter. The first experiment is a video recording of a white wall under fluorescent lightings, and the camera was fixed during the recording. In this example the nominal value of ENF is 60 Hz, and the frame rate of the camera is $\frac{1}{T} = 29.97$ fps. The intensity variations of the fluorescent lightings should follow the instantaneous energy of the AC power supply, thus exhibit a oscillation of around 120 Hz. By Equation (5), the ENF traces embedded by the row signal from the video recording should appear at around $120 + m \times 29.97$ Hz, where $m = 1, 2, 3, ...$ This matches what we observe from the spectrogram of the row signal in Figure 4.

The ENF traces can be extracted from the spectrogram of the row signals around 30 Hz, 60 Hz, 90 Hz... Compared with Figure 1, we can see that the SNR of the ENF signal is now significantly improved. We estimate the ENF signal by computing the dominant instantaneous frequency within a small range around the frequency of interest. The estimated ENF signal from this recording along with the reference ENF signal simultaneously measured from the power mains are plotted in Figure 5 (a). They are appropriately shifted to lie within the same dynamic range, as only the variation trends are of interest. The ENF signals from the video recording and the power measurement exhibit very similar trends of variations. The correlation coefficient between them as a function of the relative time lag is plotted in Figure 5 (b), and a clear peak is observed at the correct lag of 0 second.

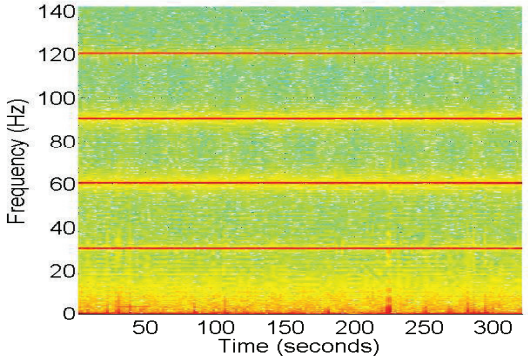
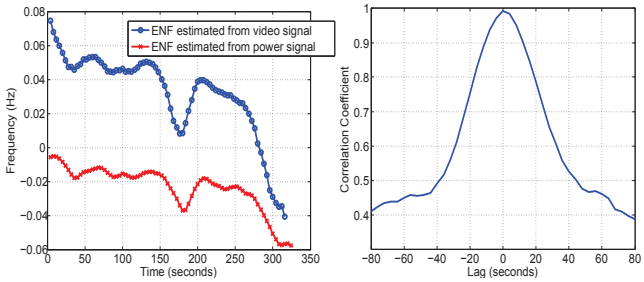


Fig. 4. The spectrogram of the row signal from a white wall video recording. Figure is best viewed in color.



(a) (b)

Fig. 5. (a) The ENF signals (appropriately shifted) extracted from a test video of white wall and the power measurement. (b) Correlation coefficient between the video and power ENF signal as a function of relative time lag.

3.2. Video with Motion

Extracting ENF traces from video recordings of scenes with moving objects is more challenging. In this case, Equation (6) does not hold any more, and the method in the previous subsection would no longer work. We explore two preliminary schemes to address this problem under different assumptions.

In the first scheme, we assume that there is no non-rigid deformation or occlusion, and there is no object entering or departing from the scene during the video recording. Under this assumption, each point in a frame can be found in other frames with a certain spatial shift. Denote by $F(r, c, n)$ the pixel at the r^{th} row and c^{th} column of frame n. It again can be considered as a combination of the visual component $V(r, c, n)$ and the ENF component $E(r, n)$:

$$F(r, c, n) = V(r, c, n) + E(r, n). \quad (9)$$

Following the constant intensity assumption from the motion estimation literature, for any $(n, \delta n)$, there is a spatial shift$(\delta r, \delta c)$ so that

$$V(r, c, n) = V(r + \delta r, c + \delta c, n + \delta n). \quad (10)$$

For a frame $F(r, c, n)$, we can find its motion-compensated version using other frames. Denoting the average of the motion-compensated frames by $\tilde{F}(r, c, n)$, given that the temporal average

of the ENF component tends to be 0, we have

$$\begin{aligned}\tilde{F}(r,c,n) &= \frac{1}{N}\sum_{i=1}^{N} F(r+\delta r_i, c+\delta c_i, n+\delta n_i) \\ &= \frac{1}{N}\sum_{i=1}^{N}\{V(r+\delta r_i, c+\delta c_i, n+\delta n_i) \\ &\quad + E(r+\delta r_i, c+\delta c_i, n+\delta n_i)\} \\ &\simeq V(r,c,n). \end{aligned} \tag{11}$$

Then we can subtract the average of the motion-compensated frames from the original frame to obtain

$$\Delta F(r,c,n) = F(r,c,n) - \tilde{F}(r,c,n) \simeq E(r,n). \tag{12}$$

The ENF signal can then be estimated from $\Delta F(r,c,n)$ using the techniques described before.

The key to the motion compensation based scheme is finding the point-wise spatial displacement between video frames. We have adopted the optical flow approach in this work, and have used the implementation by [15].

An experiment was conducted to verify the proposed scheme. The camera was held by hand to shoot a poster displayed under indoor lightings. During recording, the camera was shaken slightly so that there is noticeable global motion in the video, as can be seen in Figure 6. We used the optical flow method to obtain the spatial shift between frames, and extract the ENF signal as decried above. Figure 7 shows that the ENF signal estimated from the poster video presents a high correlation with the ENF reference from power measurements.

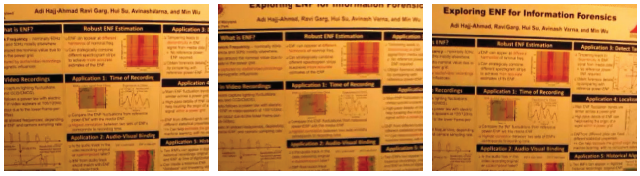
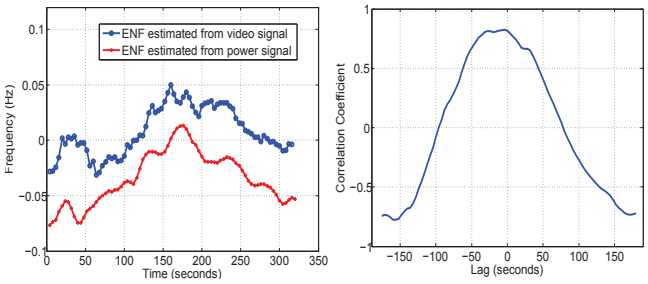


Fig. 6. Sample frames of the poster test video.



(a) (b)

Fig. 7. (a) The ENF signals (appropriately shifted) extracted from the test video of poster and the power measurement. (b) Correlation coefficient as a function of relative time lag.

The second approach to extracting ENF traces from videos with motion is to find and utilize only the regions that are static in the video frames. We demonstrate this idea with the following example. We made a video recording of a moving Hexbug toy in a room with indoor lightings. The Hexbug is a robotic toy of fast movement powered by the vibrations of a built-in battery motor. Figure 8 shows several frames of the video. For a current frame, we compare it with several neighboring frames, and find the regions in which the content has no significant change. These regions are combined to form a smoothened frame. We then take the difference of pixel values between the current frame and the smoothened frame, and use the resulting residue values to estimate the ENF signal. The ENF signals extracted from the Hexbug video and its reference ENF signal from power mains are plotted in Figure 9 (a). They exhibit similar variation trends and a correlation peak is observed when they are aligned, as seen from Figure 9 (b).

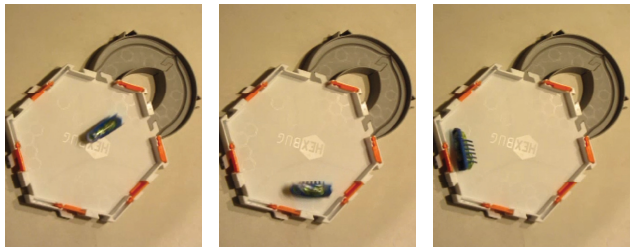
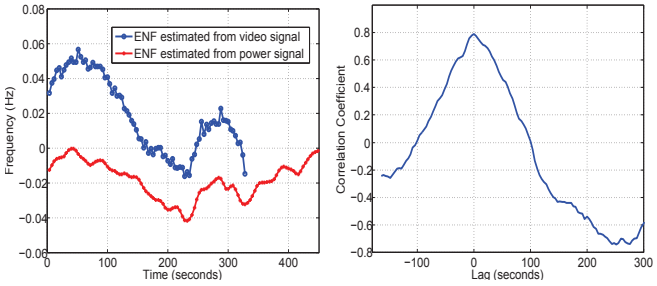


Fig. 8. Sample frames of the hexbug test video.



(a) (b)

Fig. 9. (a) The ENF signals extracted from the test video of hexbug and the power measurement (appropriately shifted). (b) Correlation coefficient as a function of relative time lag.

4. CONCLUSIONS

In summary, this work has studied the problem of extracting ENF traces from videos. We have exploited the rolling shutter of a CMOS imaging sensor and treat each line as an ENF impacted signal sample in order to compensate for the low frame rate of video recordings. The rolling shutter mechanism was analyzed using a filter bank model and multirate signal processing theory. Two methods of ENF extraction for videos with motions have been proposed. The first method is motion compensation using estimated pixel shift among video frames; and the second method is to identify and utilize static regions in the video. The experimenting results have demonstrated the effectiveness of the proposed methods.

5. REFERENCES

[1] C. Grigoras, "Applications of ENF criterion in forensics: Audio, video, computer and telecommunication analysis,"

Foresnsic Science International, vol. 167(2-3), pp. 136–145, April 2007.

[2] R. W. Sanders, "Digital authenticity using the electric network frequency," in *33rd AES International Conference on Audio Forensics, Theory and Practice*, June 2008.

[3] D. Rodriguez, J. Apolinario, and L. Biscainho, "Audio authenticity: Detecting ENF discontinuity with high precision phase analysis," *IEEE Transactions on Information Forensics and Security*, vol. 5(3), pp. 534–543, Septemper 2010.

[4] R. Garg, A. Hajj-Ahmad, and M. Wu, "Geo-location estimation from electrical network frequency signals," in *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, May 2013.

[5] A. Hajj-Ahmad, R. Garg, and M. Wu, "ENF based location classification of sensor recordings," in *IEEE Int. Workshop on Info. Forensics and Security (WIFS)*, Nov. 2013.

[6] A. Hajj-Ahmad, R. Garg, and M. Wu, "Instantaneous frequency estimation and localization for ENF signals," in *APSIPA Annual Summit and Conference*, Dec. 2012.

[7] H. Su, A. Hajj-Ahmad, M. Wu, and D. Oard, "Exploring the use of ENF for multimedia synchronization," in *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[8] A. Hajj-Ahmad, R. Garg, and M. Wu, "Spectrum combining for ENF signal estimation," *IEEE Signal Processing Letters*, vol. 20(9), 2013.

[9] R. Garg, A. Varna, and M. Wu, "'seeing' ENF: natural time stamp for digital video via optical sensing and signal processing," in *19th ACM International Conference on Multimedia*, Nov. 2011.

[10] R. Garg, A. L. Varna, A. Hajj-Ahmad, and M. Wu, "'seeing' ENF: Power signature based timestamp for digital multimedia via optical sensing and signal processing," *IEEE Trans. Info. Forensics Security*, vol. 8(9), 2013.

[11] C.-K. Liang, L.-W. Chang, and H. H. Chen, "Analysis and compensation of rolling shutter effect," *IEEE Transactions on Image Processing*, vol. 17(8), pp. 1323–1330, 2008.

[12] O. Ait-Aider, A. Bartoli, and N. Andreff, "Kinematics from lines in a single rolling shutter image," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[13] J. Gu, Y. Hitomi, T. Mitsunaga, and S. Nayar, "Coded rolling shutter photography: Flexible space-time sampling," in *IEEE International Conference on Computational Photography (ICCP)*, 2010.

[14] P. P. Vaidyanathan, *Multirate Systems And Filter Banks*, Prentice Hall, 1992.

[15] C. Liu, "Beyond pixels: Exploring new representations and applications for motion analysis," *Doctoral Thesis, Massachusetts Institute of Technology*, May 2009.