INVISIBLE GEO-LOCATION SIGNATURE IN A SINGLE IMAGE

Chau-Wai Wong[*†] *Adi Hajj-Ahmad*[⋆†] *Min Wu*[†]

[*] North Carolina State University, Raleigh, NC [⋆] GE Digital, San Ramon, CA
[†] University of Maryland, College Park, MD

ABSTRACT

Geo-tagging images of interest is increasingly important to law enforcement, national security, and journalism. Many images today do not carry location tags that are trustworthy and resilient to tampering; and the landmark-based visual clues may not be readily present in every image, especially in those taken indoors. In this paper, we exploit an invisible signature from the power grid, the Electric Network Frequency (ENF) signal, which can be inherently recorded in a sensing stream at the time of capturing and carries useful location information. It is, however, very challenging to extract an ENF signal from a single image, as compared to the recent art in extracting ENF traces from audio and video. This paper presents novel investigations toward this challenge, by synergistically exploring the rolling shutter effect of CMOS imaging sensors and entropy differences of composite signals. We study quantitatively the relationship between the ENF strength and its detectability from a single image, and bring out a unique forensics capability of invisible traces that shine a light on an image's capturing location.

***Index Terms*—** Geo-Tagging, Frequency Estimation, Electric Network Frequency (ENF), Rolling Shutter

1. INTRODUCTION

The digital era in which we are living today has seen an unprecedented amount of digital audio, image, and video being generated every day. The information they carry about the location origin where media files were captured is valuable for law enforcement, national security, and journalism. In this paper, we focus on the problem of deriving information from a single image regarding the location where it was captured, especially in challenging scenarios which existing technologies cannot sufficiently address.

A number of digital cameras today, including many cellphone cameras, can associate the built-in GPS input with the image/video's metadata field relating to the location in which the image/video was captured. This metadata, however, is not always available and can be tampered with in a relatively easy way.

Recent advances, notably through the FINDER program by the U.S. Intelligence Advanced Research Project Activity (IARPA) agency [1], utilized computer vision techniques and a big data paradigm to exploit visible terrains and landmarks appearing in an image and match them with a variety of geospatial data to identify the image's geographic location. These technologies would still face considerable challenges in cases of absence of visible landmarks, such as when the background of outdoor images does not have differentiating features, or when an image is captured indoors. The latter scenarios are common in such global efforts as fighting crime on child exploitations [2–4], where a significant portion of images of such mistreatment and abuse are taken indoors in private locations by perpetrators, and the Internet has eliminated the national borders on such crime increasing the difficulty in tracking down the perpetrators. Having new technologies to locate where such images were taken, even on a coarse level, will provide an unprecedented tool to aid global investigations, supporting law enforcement in appropriate jurisdictions.

In this work, we explore a nearly invisible location-related signature that is inherently captured in an image. This invisible signature comes from the power distribution network whose varying frequency properties over time are known as the Electric Network Frequency (ENF) signal. There is a growing amount of multimedia forensics research recently on the ENF signal, which, over time in a multimedia file, reflects the behavior of the power grid at the time and the location where a media signal is recorded. Such a relation enabled the estimation of the time, location and integrity of corresponding multimedia signals [5–9].

Due to the temporally varying nature of the ENF signature that plays a critical role in its ability to relate to time and location, ENF extraction in the literature thus far requires the hosting signal to have a temporal nature as well, e.g., an audio or a video. Two intriguing questions to push the boundary of ENF's localization capability are: Can ENF traces be found in a single image? And can such a basic attribute as the nominal ENF value be inferred at a reasonable accuracy from a single image? This paper aims at answering these questions whereby narrowing down whether an image under question comes from a 50 Hz or 60 Hz country would provide a valuable clue to law enforcement investigations and shine a light on the jurisdiction in charge. To the best of our knowledge, this is the first work toward developing a unique forensics capability to address the challenging geo-location scenarios for individual images that do not have a visible landmark or GPS tag, especially those captured indoors.

The rest of this paper is organized as follows. Section 3 presents our proposed approaches for ENF signal extraction from a single image. Section 4 shows the experiments conducted and results obtained. Section 5 concludes the paper.

2. BACKGROUND AND PRELIMINARIES

Electric Network Frequency The Electric Network Frequency (ENF) of power distribution networks has a nominal value of 60 Hz in most of the Americas, and 50 Hz in most of the rest of the world. The instantaneous frequency of the sinusoidal variation in power networks does not stay at this nominal value. Rather, it fluctuates around the nominal value due to load changes across the power grid.

The first two authors performed the work while they were with the University of Maryland, College Park. Typos corrected on Nov. 17, 2020.

Table 1. Estimated parameters for rolling-shutter images.

	iPhone 6s	iPhone 6	iPhone 5
T_{ro} (ms)	19.8	30.9	22.6
number of rows, L	3024	2448	2448
f_{row} (kHz)	152.7	117.1	108.3

The trends in the ENF variations tend to be very similar at different points in the same grid. The changing instantaneous value of the ENF over time is what we define as the *ENF signal*.

The ENF signal has been a growing subject of multimedia forensics research in recent years due to its presence in media sensing signals. If an audio or video recording is made in an area where there is electrical activity, it is likely that this recording will capture the ENF variations at the time of recording. In audio recordings, this has been attributed to the acoustic mains hum emitted from devices connected to the power mains [10]. In video recordings, this comes from the near-invisible flickering of electric lighting [11, 12].

In this paper, we show that it is possible to find ENF traces captured in single images as well when they are taken in electrically lighted settings. In particular, ENF can be extracted from images taken by the widely used cameras with complementary metal–oxide–semiconductor (CMOS) image sensors that employ a *rolling shutter*.

Rolling Shutter and Read-Out Time Unlike cameras that employ a *global shutter* that acquires the pixels of an image frame all at the same time, a camera employing rolling shutter acquires an image frame one row at a time. Although traditionally regarded as the cause of negative artifacts in images/videos, the rolling shutter has been exploited in beneficial ways for computational photography applications [13, 14]. In this work, we shall see that this sequential read-out time mechanism of the rolling shutter while capturing a single image allows the resulting image to capture samples of an incoming electric light signal at slightly different points in time, and thus obtain a short segment of ENF traces.

With rolling shutter, each row of the frame is sequentially exposed to incoming light. The amount of time during which a camera acquires the rows of an image frame, which we denote by the read-out time T_{ro}, is specific to the camera's model line and is a value that is not typically given in its user manual or specifications list.

In order to extract ENF information found in a single image, it is important to know the read-out time T_{ro} value of the camera that had produced the image under question. In our work, we have made use of a protocol inspired by the one proposed in [15, 16] to compute the T_{ro} value of a camera at hand. Table 1 lists the estimated values we computed for the cameras on the backsides of different models of iPhone devices that are used in this paper.

ENF Signal Embedding in a Single Image The rolling shutter mechanism equivalently turns rows of images of CMOS cameras into high-frequency samplers across rows. The sampling frequency of a row can be expressed by $f_{\text{row}} = L \times T_{\text{ro}}^{-1}$, where L is the number of rows of the image. A CMOS camera captures ENF traces by recording the electric light signal. The electric light intensity embedded in the resulting image relates to the supplied electric current via a power law thus making its nominal frequency twice the nominal ENF value, i.e., 120 Hz in most of the Americas and 100 Hz in most other parts of the world. We can express parametrically the ENF signal captured in an image by considering the acquisition at different row indices i to be samples of a fluctuating signal as follows:

$$e(i) = A\cos\left(2\pi\frac{f_{\text{ENF}}}{f_{\text{row}}}\cdot i + \phi\right), \tag{1}$$

where f_{ENF} is the fluctuation frequency of the light intensity, A is the magnitude, i is the row index of the image in $\{0, \cdots, L-1\}$, and ϕ is the initial phase of ENF signal at time the first row is acquired.

When an image is taken through the rolling shutter mechanism, the light intensity signal is modulated with a native image $x(i, j)$. Although this modulation can be nonlinear, a large part can be considered as an additive component $e(i)$ to the native image content. Hence, the ENF-containing image $y(i, j)$ can be approximately modeled by adding a parametric surface to the native image content, namely, $y(i, j) = x(i, j) + e(i)$.

3. ENTROPY MINIMIZATION FOR PARAMETRIC-SURFACE–REMOVED IMAGES

As a first step toward the challenging geo-tagging problem, we address the following research question: *Can we differentiate whether a captured image contains 100 or 120 Hz ENF signal? If yes, what is the lowest level of the intrinsically embedded signal that can be detected?* To answer this question, we propose a method that will exploit i) the prior knowledge that the embedding signal has a parametric form, and ii) a largely predictable change in the statistical behavior of the image before and after ENF embedding.

Entropy Impact by ENF Embedding We observe that smooth image regions tend to become busier or have higher entropy after being corrupted by real-valued sinusoid signals. In case of an image column with constant intensity, the sole bin of the histogram before corruption starts to split into multiple bins due to the positive and negative additive values contributed by the sinusoid signal. Such corruption increases the entropy of the constant signal from zero to a positive value. In case of an image column with linearly increasing intensity values, similar statistical behavior exists even though the increase is not as drastically as the constant intensity case. The histogram of the linearly increasing intensity case before corruption is a square window. After corruption, the bins at two sides of the window will split, which may increase the entropy.

We verify the statistical behavior claimed above via simulation. We simulated the ENF embedding process by adding real-valued sinusoid signals at 50 or 60 Hz with ten randomly picked phases to columns of images. We calculated the estimated entropy increase for each column and draw the histogram of entropy increase in Fig. 1 (a). It reveals that 99% of all columns have an increase in entropy after being corrupted by sinusoid signals, with the understanding that entropy values of texture regions of images may not be sensitive to the additive sinusoid signals.

Entropy Minimization for Frequency Estimation Below we lay out the idea of estimating the frequency from an ENF corrupted image column. The statistical behavior that entropy increases after ENF corruption guarantees that if the true ENF signal is removed, the entropy will return to the entropy of the uncorrupted image content. This entropy value is a local minimum or even a global minimum of a cost function, $J_k(\mathbf{p})$, defined as the entropy of a real-valued sinusoid subtracted image column with search parameters $\mathbf{p} = [A, f, \phi]^T$. In case the cost function is in addition smooth within a small neighborhood of the true parameter

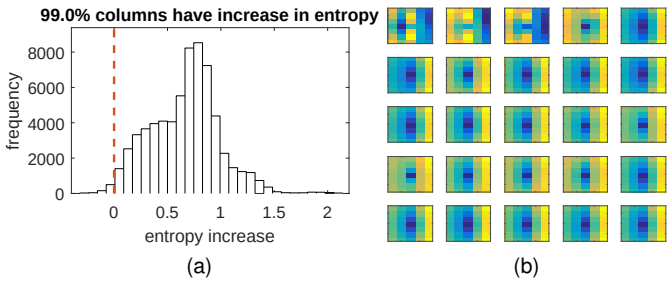

Fig. 1. (a) Histogram of column's entropy increase due to ENF signal embedding. (b) Error surfaces for 25 equally-spaced columns of an image. Darker color means lower error. Objective function parameters: amplitude, A, in x-axis; and initial phase, ϕ, in y-axis.

$\mathbf{p}_{\text{g}} = [A_g, f_g, \phi_g]^T$, the true frequency can be obtained by a parameter search.

To estimate the embedding ENF frequency from an image, we define a cost function, $\sum_{k=1}^{K} J_k(\mathbf{p})$, which is the total entropy of K-selected columns of the sinusoid-surface–removed image with the search parameter $\mathbf{p}$. As is shown in the proofs from a supplemental document (hosted on the authors' webpages), a cost function involving K columns can boost the accuracy of the estimation by at most $\sqrt{K}$ times measured in the standard deviation of the frequency estimator.

We can further enforce a strong prior knowledge that f is either 100 or 120 Hz, so that the parameter search is restricted to a 2-d space (A, ϕ). Fig. 1(b) of error surfaces for 25 far apart columns of an image reveals that, when the frequency is a correct guess, the cost functions are generally well-behaved, i.e., convex, and with unbiased minimum positions.

Proposed Algorithm We briefly outline the frequency decision task as follows. Given an image corrupted by one of the two possible ENF signals, we estimate (A, ϕ) separately for $f = 100$ and 120 Hz. When the estimation is carried out for the correct frequency, the residue is roughly the native image content; when the estimation is carried out for incorrect frequency, the residue is a combination of the native image content, and the sum of two sinusoid signals with different frequencies. Given that the native image content usually has lower entropy than the content with additional sinusoid components, we choose the candidate frequency leading to the lower entropy as our decided frequency. The pseudo code for the proposed entropy minimization method is shown in Algorithm 1.

Algorithm 1: The proposed entropy minimization method for images with embedded ENF traces.

input : An RGB image with embedded ENF at 100/120 Hz
output: $\hat{f}_{\text{ENF}} \in \mathcal{F} = \{100, 120\}$ Hz

Step 1. Initialization

Convert the image to gray-scale
Detrend image's overall intensity with RANSAC algorithm
Randomly select K gray-level columns, and augment them into a sample matrix $\mathbf{X} = [x^{(1)}, \cdots, x^{(K)}]$

Step 2. Optimization

$br(\ell) \leftarrow \min_{A,\phi} CostFunc\left(\mathbf{X}, \mathcal{F}(\ell), A, \phi\right),\ \ell = 1, 2.$

Step 3. Decision

Decide $\hat{f}_{\text{ENF}}$ between 100 Hz and 120 Hz based on the setting that gives a smaller bitrate br

Subroutine $bitrate \leftarrow CostFunc\left(\mathbf{X}, f, A, \phi\right)$

$e(i) \leftarrow A\cos(2\pi \frac{f}{f_{\text{row}}} \cdot i + \phi),\ \ \forall i$
$\hat{\mathbf{X}}(i,k) \leftarrow \mathbf{X}(i,k) - e(i),\ \ \forall (i,k)$
$bitrate\_array(k) \leftarrow entropy(hist(\hat{\mathbf{X}}(:,k))),\ \ \forall k$
return $mean(bitrate\_array)$

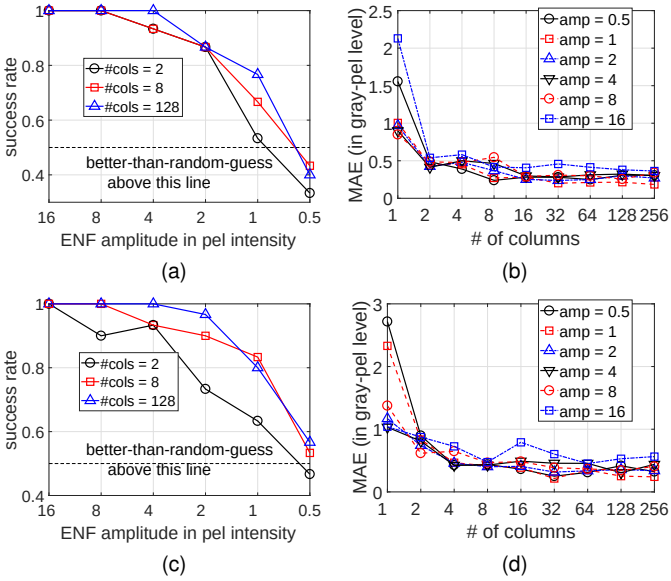

Fig. 2. Success rates for deciding images with (a) 60 Hz and (b) 50 Hz ENF as a function of the ENF strength and the number of columns used in the cost function. (b)(d) The corresponding MAEs for the amplitude estimator.

4. EXPERIMENTAL RESULTS

Images with Synthetic ENF Traces We created a set of test images to emulate those captured in 50 Hz and 60 Hz regions using six images that do not contain detectable ENF traces. Sinusoid signals with random initial phases and a list of decreasing amplitude levels, $[16, 8, 4, 2, 1, 0.5]$ in the same unit of pixel intensity with 256-level shades of gray were added to the native images to generate a total of 72 $(= 6 \times 2 \times 6)$ images. For each image and each tested number of columns used, the proposed algorithm of stochastic nature was applied five times to better reveal the performance of the algorithm on the particular image. Performances for 50 Hz and 60 Hz images were assessed separately.

The estimated probability of success decision (*aka* the success rates) were shown in Figs. 2 (a) and (c) as a function of decreasing ENF amplitude for 60 Hz and 50 Hz images, respectively. It is revealed that the success rate for both cases can be as high as 0.9 for ENF amplitude equals to 4. As the amplitude decreases further, the success rate drops to random guesses when the amplitude is between 0.5 and 1 for 60 Hz images, and is around 0.5 for 50 Hz images.

Examining the accuracy of the estimators for the nuisance parameters A and ϕ can hint at the reliability of the proposed entropy minimization method. For example, Figs. 2 (b) and (d) shows that as the number of columns used increases, the mean absolute error (MAE) generally decreases. Both plots for 50 Hz and 60 Hz images show that increasing from using one column to two columns leads

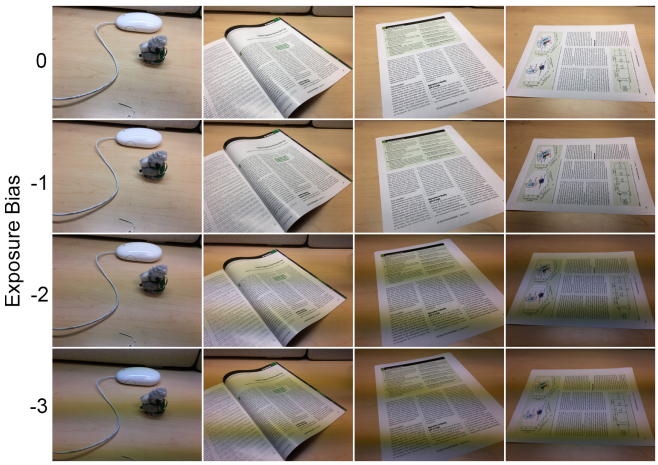

Fig. 3. Four scenes from images taken in the US with nominal ENF at 60 Hz using iPhone 6s.

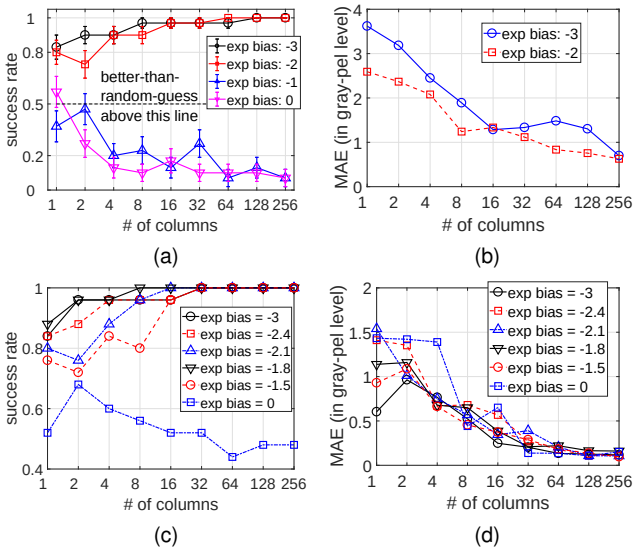

Fig. 4. Success rates for deciding images with (a) 60 Hz and (c) 50 Hz ENF traces as a function of the ENF strength and the number of columns used in the cost function. (b)(d) The corresponding MAEs for the amplitude estimators.

to the largest improvement in accuracy, and a further increase in the number of columns used leads to marginal improvement.

Real Rolling Shutter Images with ENF Traces We created a real-world image dataset using multiple smartphone cameras to take images in indoor environments. For each scene, different levels of *exposure bias* were set on the camera using each camera's built-in functionality to control the exposure time of the captured images. We acquired in the United States a dataset with 60 Hz ENF signal containing six scenes, each scene with exposure bias of $[0, -1, -2, -3]$; and in China a dataset with 50 Hz ENF signal containing five scenes, each scene with exposure bias of $[0, -1.5, -1.8, -2.1, -2.4, -3]$. A less negative exposure bias corresponds to a longer exposure time and a brighter image, and vice versa. Fig. 3 shows four scenes from the United States where the brightness of the electric lighting varies at 120 Hz; for each scene, four levels of exposure bias were shown. As we can see, a short exposure time can quite clearly capture the fluctuation of the ENF light signal. As the exposure time increases, the amplitude of the fluctuation of the captured signal decreases due to an effective convolution of a rectangular window with the electric light signal.

Figs. 4 (a)–(b) show the performance of the proposed entropy minimization on real-world camera images captured in a 60 Hz ENF country. One can observe that compared to using two columns to obtain a strong detection performance as in the synthetic case, 16 columns are needed in the real-world camera image case. It is also revealed that as the level of exposure bias reduces from -2 to -1, the method breaks as the success rate becomes dominated by being biased toward deciding 50 Hz. Combining the two aspects of the results, we anticipate that the current algorithm may start to fail on real-world camera images when the amplitude is below 10.

Figs. 4 (c)–(d) for real-world camera images captured in a 50 Hz country similarly shows that as the number of columns used increase, the success rate and estimation accuracy improves. An interesting observation is that no matter how weak the ENF signal is, the success rate for 50 Hz images is always better than random guesses. This suggests that even though the proposed entropy minimization method was not designed to be biased toward low frequency, and the experimental results for images with synthetic ENF traces do not show such bias, the bias does exist when the entropy minimization method is applied on the real-world camera images. Further investigations can be conducted to examine the impact of the modeling mismatch, e.g., when the ENF signal captured in the image is not a perfect sinusoid.

Discussions In addition to the additive model [11, 12, 17, 18] that is resilient to the global intensity bias used in this paper, we also tried the multiplicative model that is resilient to albedo and illumination changes. Our theoretical analysis showed that if the global intensity bias of the camera captured and processed image is negligible compared to the intensity of the content, the multiplicative model would be precise. Our experimental results showed that an algorithm developed based on the additive model has a better performance on real-world data than the multiplicative model.

It is important to recall the significant challenge in extracting ENF signals from single images, as it is extremely difficult to overcome the dominating visual content and reveal the weak ENF signals. In contrast, the ENF extraction from video [11, 17] can take advantage of multiple frames to estimate the native visual content and separate the ENF from the visual content. Despite the difficulty, the results in this section have shown promising capability by our proposed algorithms as the first attempt to tackle the challenge.

5. CONCLUSION

In this paper, we explored the use of an invisible power signature, the ENF signal, that may be embedded in images acquired by CMOS cameras for geo-tagging purposes. Specifically, we addressed a very challenging research question of whether the basic attributes of an embedded ENF signal can be correctly identified. We proposed the entropy minimization method for parametric-surface–removed images. Experimental results show that method is able to make high accuracy decisions when the ENF traces are strong. In this proof-of-concept work, we have demonstrated a unique forensics capability of extracting invisible traces to help narrow down the capturing geographic region of an image. The next step would be to examine and improve the performance on a larger scale investigation with more cameras and images, and explore additional location information that can be inferred from the ENF traces in an image.

6. REFERENCES

[1] "Finder (IARPA Research Program)," [Online]. Available: https://www.iarpa.gov/index.php/research-programs/finder, Accessed November 2016.

[2] "Operation Predator – Wikipedia," [Online]. Available: https://en.wikipedia.org/wiki/Operation_Predator, Accessed November 2016.

[3] "Child Exploitation Investigations Unit, U.S. Department of Homeland Security," [Online]. Available: https://www.ice.gov/predator, Accessed November 2016.

[4] "Child Protection from Violence, Exploitation and Abuse, UNICEF at United Nation," [Online]. Available: https://www.unicef.org/protection/, Accessed November 2016.

[5] C. Grigoras, "Digital audio recordings analysis: The electric network frequency (ENF) criterion," *International Journal of Speech, Language and the Law*, vol. 12, no. 1, pp. 63–76, 2005.

[6] ——, "Applications of ENF analysis in forensic authentication of digital audio and video recordings," *Journal of the Audio Engineering Society*, vol. 57, no. 9, pp. 643–661, 2009.

[7] D. P. N. Rodríguez, J. A. Apolinário, and L. W. P. Biscainho, "Audio authenticity: Detecting ENF discontinuity with high precision phase analysis," *IEEE Transanctions on Information Forensics and Security*, vol. 5, no. 3, pp. 534–543, Sep. 2010.

[8] P. A. Andrade Esquef, J. A. Apolinário, and L. W. P. Biscainho, "Edit detection in speech recordings via instantaneous electric network frequency variations," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2314–2326, Dec. 2014.

[9] A. Hajj-Ahmad, R. Garg, and M. Wu, "ENF-based region-of-recording identification for media signals," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 6, pp. 1125–1136, Jun. 2015.

[10] N. Fechner and M. Kirchner, "The humming hum: Background noise as a carrier of ENF artifacts in mobile device audio recordings," in *Proceedings of the International Conference on IT Security, Incident Management, and IT Forensics*, May 2014, pp. 3–13.

[11] R. Garg, A. L. Varna, and M. Wu, "'Seeing' ENF: Natural time stamp for digital video via optical sensing and signal processing," in *Proceedings of the ACM International Conference on Multimedia*, 2011, pp. 23–32.

[12] 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 Transactions on Information Forensics and Security*, vol. 8, no. 9, pp. 1417–1432, Sep. 2013.

[13] O. Ait-Aider, A. Bartoli, and N. Andreff, "Kinematics from lines in a single rolling shutter image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2007, pp. 1–6.

[14] J. Gu, Y. Hitomi, T. Mitsunaga, and S. Nayar, "Coded rolling shutter photography: Flexible space-time sampling," in *Proceedings of the IEEE International Conference on Computational Photography*, Mar. 2010, pp. 1–8.

[15] A. Hajj-Ahmad, S. Baudry, B. Chupeau, and G. Doërr, "Flicker forensics for pirate device identification," in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, ser. IH&MMSec, Portland, Oregon, USA, 2015, pp. 75–84.

[16] A. Hajj-Ahmad, S. Baudry, B. Chupeau, G. Doërr, and M. Wu, "Flicker forensics for camcorder piracy," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 1, pp. 89–100, Jan. 2017.

[17] H. Su, A. Hajj-Ahmad, R. Garg, and M. Wu, "Exploiting rolling shutter for ENF signal extraction from video," in *Proceedings of the IEEE International Conference on Image Processing*, Oct. 2014, pp. 5367–5371.

[18] H. Su, A. Hajj-Ahmad, C.-W. Wong, R. Garg, and M. Wu, "ENF signal induced by power grid: A new modality for video synchronization," in *Proceedings of the ACM International Workshop on Immersive Media Experiences*, Nov. 2014, pp. 13–18.

Note: This is a supplementary file for paper "Invisible geo-location signature in a single image," published in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing* by Chau-Wai Wong, Adi Hajj-Ahmad, and Min Wu.

APPENDIX

Using single column to estimate parameters

Let us define a column of smooth native image content as $\mathbf{x}^{(k)} \in \mathbb{R}^{N\times 1}$, and an embedding ENF signal as $\mathbf{e}$, where

$$e(i) = A_{\mathrm{g}} \cos\left(2\pi \frac{f_{\mathrm{g}}}{f_{\mathrm{row}}} \cdot i + \phi_{\mathrm{g}}\right). \tag{2}$$

We define $\mathbf{p}_{\mathrm{g}} = [A_{\mathrm{g}}, f_{\mathrm{g}}, \phi_{\mathrm{g}}]^T$ to be a vector of the ground truth parameters. The corrupted sensing signal is therefore $\mathbf{y} = \mathbf{x}^{(k)} + \mathbf{e}$. We set up a cost function to estimate $\mathbf{p}$ as follows:

$$J_k(\mathbf{p}) = \text{entropy}\left(\text{hist}\left[\left\{x^{(k)}(i) + r(i;\mathbf{p}_{\mathrm{g}},\mathbf{p})\right\}_{i=1}^{N}\right]\right) \tag{3a}$$

$$= \text{entropy}\left(\text{hist}\left[\mathbf{x}^{(k)} + \mathbf{r}(\mathbf{p}_{\mathrm{g}},\mathbf{p})\right]\right), \tag{3b}$$

where $r(i;\mathbf{p}_{\mathrm{g}},\mathbf{p}) = e(i) - A\cos\left(2\pi\frac{f}{f_{\mathrm{row}}}\cdot i + \phi\right)$. We denote $\hat{\mathbf{p}}(\mathbf{x}^{(k)})$ as a numerical suboptimal solution to the optimization problem

$$\min_{\mathbf{p}} J_k(\mathbf{p}). \tag{4}$$

By taking into consideration the randomness of the image, the suboptimality of the numerical algorithm for finding the optimal solution, we model $\hat{\mathbf{p}}(\mathbf{x}^{(k)})$ as a random vector with a small bias $\mathbf{b}$ and a positive definite variance-covariance matrix $\mathbf{\Sigma}$, namely,

$$\mathbb{E}\left[\hat{\mathbf{p}}(\mathbf{x}^{(k)})\right] = \mathbf{p}_{\mathrm{g}} + \mathbf{b}, \quad \text{VarCov}\left(\hat{\mathbf{p}}(\mathbf{x}^{(k)})\right) = \mathbf{\Sigma}. \tag{5}$$

Using multiple columns to improve the accuracy of the estimation

We use a random sample $(\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(K)})$ to improve the accuracy of the estimation. We use $\hat{\mathbf{p}}\left(\{\mathbf{x}^{(k)}\}_{k=1}^{K}\right)$ to denote a numerical suboptimal solution to the multi-column optimization problem

$$\min_{\mathbf{p}} \sum_{k=1}^{K} J_k(\mathbf{p}). \tag{6}$$

We conduct Taylor-series expansion around $\hat{\mathbf{p}}(\mathbf{x}^{(k)})$ for $J_k(p)$, and obtain

$$J_k(\mathbf{p}) = J_k\left(\hat{\mathbf{p}}(\mathbf{x}^{(k)})\right) + \left[\mathbf{p} - \hat{\mathbf{p}}(\mathbf{x}^{(k)})\right]^T \nabla J_k\left(\hat{\mathbf{p}}(\mathbf{x}^{(k)})\right) + \frac{1}{2}\left[\mathbf{p} - \hat{\mathbf{p}}(\mathbf{x}^{(k)})\right]^T \nabla^2 J_k \left[\mathbf{p} - \hat{\mathbf{p}}(\mathbf{x}^{(k)})\right], \tag{7}$$

where $\nabla^2 J_k$ is a Hessian matrix whose individual components are 2nd-order partial derivatives evaluated near $\hat{\mathbf{p}}(\mathbf{x}^{(k)})$. We further assume $J_k(\cdot)$ is continuous hence $\nabla^2 J_k$ is a symmetric matrix.

Substituting the Taylor expanded cost function for each column in (7) into the multi-column cost function in (6), and by setting the gradient with respect to $\mathbf{p}$ to zero, we obtain the optimal analytical solution to (6) as follows:

$$\hat{\mathbf{p}}\left(\{\mathbf{x}^{(k)}\}_{k=1}^{K}\right) = \left(\sum_{k=1}^{K}\nabla^2 J_k\right)^{-1}\left\{\sum_{k=1}^{K}\left[\nabla^2 J_k \hat{\mathbf{p}}(\mathbf{x}^{(k)}) - \nabla J_k\left(\hat{\mathbf{p}}(\mathbf{x}^{(k)})\right)\right]\right\} \tag{8a}$$

$$= \sum_{k=1}^{K}\mathbf{W}_k\,\hat{\mathbf{p}}(\mathbf{x}^{(k)}) - \left(\sum_{k=1}^{K}\nabla^2 J_k\right)^{-1}\left(\sum_{k=1}^{K}\nabla J_k\right) \tag{8b}$$

$$\approx \sum_{k=1}^{K}\mathbf{W}_k\,\hat{\mathbf{p}}(\mathbf{x}^{(k)}), \tag{8c}$$

where $\mathbf{W}_k = \left(\sum_{k=1}^{K}\nabla^2 J_k\right)^{-1}\nabla^2 J_k$. The sample average of gradients in (8b) converges to $\mathbf{0}$ in distribution as K increases. Hence,

$$\mathbb{E}\left[\hat{\mathbf{p}}\left(\{\mathbf{x}^{(k)}\}_{k=1}^{K}\right)\right] \approx \mathbf{p}_{\mathrm{g}} + \mathbf{b}, \tag{9}$$

$$\text{VarCov}\left(\hat{\mathbf{p}}\left(\{\mathbf{x}^{(k)}\}_{k=1}^{K}\right)\right) \approx \sum_{k=1}^{K}\mathbf{W}_k\,\mathbf{\Sigma}\,\mathbf{W}_k^T. \tag{10}$$

The above results reveal that by using K columns, the estimator has a significantly reduced variance and a similar bias as the single column case.

In the 1-d case, the variance formula is degenerated to $\text{Var}(\hat{p}) \approx \left(\sum_{k=1}^{K} w_k^2\right)\sigma^2$. When the weight vector is uniform, i.e., $w_i = \frac{1}{K}$, the best variance reduction is achived, i.e., $\text{Var}(\hat{p}) \approx \sigma^2/K$.