T.C.
BURSA ULUDAĞ UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

MODERN TECHNIQUES IN FORENSIC ANALYSIS OF MULTIMEDIA SIGNALS

Saffet VATANSEVER
ORCID: 0000-0002-4680-1263

Assoc. Prof. Dr. Ahmet Emir DİRİK
ORCID: 0000-0002-9174-0367

(Supervisor)

PhD THESIS

DEPARTMENT OF ELECTRONICS ENGINEERING

BURSA – 2019

THESIS APPROVAL

This thesis titled "MODERN TECHNIQUES IN FORENSIC ANALYSIS OF MULTIMEDIA SIGNALS" and prepared by Saffet VATANSEVER has been accepted as a **PhD THESIS** in Bursa Uludağ University Graduate School of Natural and Applied Sciences, Department of Electronics Engineering following a unanimous vote of the jury below.

Supervisor : Assoc. Prof. Dr. Ahmet Emir Dirik

Head : Assoc. Prof. Dr. Ahmet Emir Dirik — Signature
ORCID: 0000-0002-6200-1717
Bursa Uludağ University,
Faculty of Engineering,
Department of Computer Engineering

Member: Assoc. Prof. Dr. Figen Ertas — Signature
ORCID: orcid.org/0000-0003-4868-8425
Bursa Uludağ University,
Faculty of Engineering,
Department of Electrical and Electronics Engineering

Member: Assoc. Prof. Dr. Fatih Çavdur — Signature
ORCID: 0000-0001-8054-5606
Bursa Uludağ University,
Faculty of Engineering,
Department of Industrial Engineering

Member: Assoc. Prof. Dr. Hakan Gürkan — Signature
ORCID: 0000-0002-7008-4778
Bursa Technical University,
Faculty of Engineering and Natural Sciences,
Department of Electrical and Electronics Engineering

Member: Assoc. Prof. Dr. Ahmet Mert — Signature
ORCID: 0000-0003-4236-3646
Bursa Technical University,
Faculty of Engineering and Natural Sciences,
Department of Mechatronics Engineering

I approve the above result

Prof. Dr. Hüseyin Aksel EREN
Institute Director
13/12/2019

I declare that this thesis has been written in accordance with the following thesis writing rules of the U.U Graduate School of Natural and Applied Sciences;

- All the information and documents in the thesis are based on academic rules,
- audio, visual and written information and results are in accordance with scientific code of ethics,
- in the case that the works of others are used, I have provided attribution in accordance with the scientific norms,
- I have included all attributed sources as references,
- I have not tampered with the data used,
- and that I do not present any part of this thesis as another thesis work at this university or any other university.

13./12/2013.

Saffet VATANSEVER

ÖZET

Doktora Tezi

ÇOKLU ORTAM SİNYALLERİNİN ADLİ KANIT ANALİZİNDE MODERN TEKNİKLER

Saffet VATANSEVER

Bursa Uludağ Üniversitesi
Fen Bilimleri Enstitüsü
Elektronik Mühendisliği Anabilim Dalı

Danışman: Doç. Dr. Ahmet Emir DİRİK

Günümüz bilgi çağında, dijital kayıtların çok kolay bir şekilde manipüle edilebilir ya da yapay olarak oluşturulabilir olması, dijital kayıt dosyalarının adli analizini önemli bir araştırma konusu haline getirmiştir. Bu kapsamda, ENF (Electric Network Frequency - elektrik şebeke frekansı) ve PRNU (Photo Response Non Uniformity - ışığa olan emsalsiz sensör tepkisi) en çok ilgilenilen araştırma konuları arasındadır. ENF, elektrik şebekesi geriliminin frekansı olup, talep edilen ve harcanan güç arasındaki dengesizliğe bağlı olarak zaman içinde sürekli dalgalanmalar yapar. ENF, belirli koşullarda ses ve video kayıtlarına istemsiz olarak eklenmekte olup, bu kayıtlardan ENF sinyalinin kestirimi yapılabilir. ENF, kayıt süresi doğrulama, içerik doğrulama, videolarda ses ve görüntü senkronizasyonu, elektrik şebekesi tanımlama gibi çeşitli adli analiz işlemlerinde kullanılabilir. PRNU, her sensörün ışığa karşı olan emsalsiz tepkisi olup sensör üretiminde kullanılan silikon levhanın homojen olmayan yapısından ve üretim sürecindeki önlenemeyen kusurlardan meydana gelir. Aynı levhadan üretilen aynı marka ve model kamera sensörlerinin bile PRNU bileşenlerinin farklı olması, PRNU'nun emsalsiz sensör parmak izi olarak değerlendirilmesini sağlar. PRNU, çekilen her imge ve video çerçevesine istemsiz olarak eklenmekte olup, bu görüntülerden PRNU gürültüsünün kestirimi yapılabilir. PRNU temel olarak imge ve videoların kaydedildiği kaynak cihazın tanımlanmasında kullanılır. Bu tez çalışması kapsamında, ses ve video dosyalarının ENF tabanlı adli analizi üzerine kapsamlı bir çalışma yapılmış olup, videolarda ENF varlık/yokluk tespiti, "rolling shutter" mekanizması ile örneklenen videoların her bir çerçevesindeki bekleme süresine bağlı olarak ENF gücünün ve temel ENF harmoniğinin frekansının nasıl değiştiği, video çerçevelerindeki bu bekleme süresinin kestirimi, yine "rolling shutter" mekanizması ile örneklenen videoların kayıt zamanının doğrulanması gibi bir dizi yeni yöntem geliştirilmiş ve sunulmuştur. Bu tezde ayrıca, sorgulanan sosyal medya video çiftlerinin aynı kamera ile kaydedilip kaydedilmediği konusunda PRNU tabanlı karşılaştırmalı bir analiz sunulmuştur.

Anahtar Kelimeler: ENF, elektrik şebeke frekansı, adli kanıt analizi, adli bilişim, rolling shutter, kamera doğrulama, akustik şebeke gürültüsü, video kayıt zamanı doğrulama, PRNU, sensör gürültüsü, SPN, PRNU, kaynak kamera tespiti, kamera tanılama.
2019, x + 91 sayfa.

ABSTRACT

PhD Thesis

MODERN TECHNIQUES IN FORENSIC ANALYSIS OF MULTIMEDIA SIGNALS

Saffet VATANSEVER

Bursa Uludağ University
Graduate School of Natural and Applied Sciences
Department of Electronics Engineering

Supervisor: Assoc. Prof. Dr. Ahmet Emir DİRİK

Media forensics has become a research field of great importance in the information age as digital recordings can straightforwardly be edited, manipulated, or artificially created. The forensic criterias ENF (Electric Network Frequency) and PRNU (Photo Response Non Uniformity) are among the most interested research fields. ENF is voltage frequency of mains electricity, and it shows a continuous fluctuation in time depending on the gap between demanded and supplied power. ENF is intrinsically integrated into audio and video recordings in certain conditions, and it can be estimated from these recordings. ENF can be used for various forensic tasks including time-of-recording verification, media authentication, multimedia synchronization and power grid identification. PRNU is distinct response of each photo-sensor to light caused by non-homogeneous structure of silicon wafers and unavoidable defects in the manufacturing process. This unique characteristic of each camera sensor, even of those produced from the same wafer, and hence of the same brand and model, makes PRNU to be treated as a sensor fingerprint. PRNU noise is inherently integrated into each exposed image or video frame and can be estimated from them. PRNU can be used mainly for identification of the source device by which a given image or video is recorded. In this thesis, we provide a comprehensive study on ENF based audio and video forensics and present a number of novel methods including: proposal of ENF presence detection algorithm for video; creation of a model for where the frequency of the primary video ENF harmonic is shifted, and how the captured ENF's power is attenuated owing to the idle period in rolling shutter mechanism; development of a new method for idle period extraction; exploration of a new technique for time-of-recording verification task for rolling shutter exposed videos. We also provide a comparative analysis on PRNU based source camera attribution for social media video pairs.

Key words: electric network frequency, ENF, multimedia forensics, video forensics, camera forensics, camera verification, idle period, rolling shutter, mains hum, time-stamp verification, time-of-recording, sensor pattern noise, SPN, PRNU, camera identification, camera attribution.
2019, x + 91 pages.

ACKNOWLEGDEMENT

First and foremost, I would like to thank my beloved family; my wife, my brother and my parents for giving me an unconditional and constant love, support, encouragement and patience. They highly impacted my life and made me who I am today. Their role in pursuing my ambitions and dreams is beyond words. I dedicate this dissertation to them and my newborn baby.

I would like to express my sincere gratitude to my supervisor Assoc. Prof. Dr. Emir Dirik for his invaluable support in continuous guidance, advice, understanding, and patience. He has been a mentor to me in every aspect throughout the journey of my PhD study. I have learned too much from him. He provided an outstanding contribution in my professional career and has never hesitated to help me. He always encouraged me critical thinking, comprehensive research and outstanding presentation. I always felt lucky to work with him.

Special thanks to Prof. Nasir Memon for his enlightening expertise, outstanding guidance and insightful comments in the scope of the collaborative work "Developing Novel Video Forensics Tools and Methods Utilizing Electrical Network Frequency" as part of Medifor project. I feel I am privileged for having known him and to have been working with him.

I would like to thank Emmanuel Kiegaing Kouokam for providing me with his source code for block-based PRNU noise estimation from video and for his generous assistance in the relevant technique, which saved me a substantial amount of time.

I would like to thank Assoc. Prof. Dr. Hüsrev Taha Sencar for his support and constructive feedback; and the other members of my thesis committee, Assoc. Prof. Dr. Figen Ertaş,, Assoc. Prof. Hakan Gürkan, Assoc. Prof. Ahmet Mert and Assoc. Prof. Dr. Fatih Çavdur for their valuable comments.

I would like to thank Presidency of Defense Industry of Turkey for providing me a privilege for working as a researcher in the project titled "PRNU Sensor Noise based Source Device Identification for Video Files" as part of SAYP (Researcher Development Program for Defense Industry). I also would like to thank Dr. Aykut KOÇ for the collaboration we had during this project.

I would like to thank TÜBITAK (The Scientific and Technological Research Council of Turkey) for the grant they provided me as part of 2211-A General Domestic PhD Scholarship Program. I would also like to thank Ministry of National Education of Turkey for the grant they provided me during MSc as part of YLSY scholarship, which paved me the way for my academic career.

Saffet VATANSEVER
…/…/…….

CONTENTS

Page
ÖZET i
ABSTRACT ii
ACKNOWLEDGEMENT iii
SYMBOLS AND ABBREVIATIONS v
FIGURES vi
TABLES ix
1. INTRODUCTION 1
1.1. Motivation 1
1.2. ENF (Electric Network Frequency) Forensics 2
1.3. PRNU (Photo Response Non-Uniformity) Forensics 4
1.4. Main Contributions and Dissertation Organization 7
2. THEORETICAL BASICS AND LITERATURE REVIEW 11
2.1. ENF-based Media Forensics 11
2.1.1. ENF power model 11
2.1.2. ENF in audio 12
2.1.3. ENF estimation for audio 13
2.1.4. Ground-truth ENF acquisition 14
2.1.5. Light source flicker and ENF in video 15
2.1.6. ENF estimation for video 15
2.1.7. Idle period effect on ENF 17
2.1.8. Idle period estimation 20
2.2. PRNU-based Media Forensics 21
2.2.1. Pattern noise of imaging sensors 21
2.2.2. PRNU estimation for image 22
2.2.3. PRNU estimation for video 23
2.2.4. PRNU-based source camera identification 25
3. MATERIALS AND METHODS 27
3.1. Superpixel-based ENF Estimation for Video 27
3.2. Detecting the Presence of ENF in Video 29
3.3. An Analytical Model for ENF Dependence on Idle Period 31
3.4. Proposed Idle Period Estimation Approach 35
3.5. Improved ENF-based Video Time-of-Recording Verification Method 37
4. RESULTS AND DISCUSSION 40
4.1. Experimental Work with ENF-based Media Forensics 40
4.1.1. A study on Turkey's electricity system 40
4.1.2. A comparative work of sources of ENF in audio and of microphone types 40
4.1.3. Evaluation of proposed superpixel-based ENF presence detector 46
4.1.4. Evaluation of proposed idle period estimation approach 49
4.1.5. Experiments with improved time-of-recording verification 64
4.1.6. Factors affecting ENF forensics in video 66
4.2. PRNU-based Source Camera Attribution for Social Media Video Pairs 75
4.2.1. Experimental work on YouTube videos 77
4.2.2. Experimental work on WhatsApp videos 81
5. CONCLUSION 84
REFERENCES 87
RESUME 91

SYMBOLS and ABBREVIATIONS

Symbols	Definition
$\mathbf{I}$	An image or a video frame
$\mathbf{F}$	Reference PRNU estimate
$\mathbf{M}^{(i)}$	Mask of *i* th frame
$\mathbf{N}$	PRNU noise of an image or a video frame
$\mathbf{K}$	PRNU noise of a video

Abbreviation	Definition
AUC	Area Under Curve
CCD	Charged Coupled Device
CFL	Compact Fluorescent Lamp
CFA	Color Filter Array
CMOS	Complementary Metal Oxide Semiconductor
DTFT	Discrete Time Fourier Transform
LED	Light Emitting Diode
NCC	Normalized Cross-Correlation
PCE	Peak-to-Correlation Energy
PRNU	Photo-Response Non-Uniformity
ROC	Receiver Operating Characteristic
SPN	Sensor Pattern Noise
STFT	Short Time Fourier Transform

FIGURES

Page

Figure 2.1. Overall ENF extraction procedure from audio 13
Figure 2.2. The intermediate circuit for extraction of ground-truth ENF directly from mains power outlet via sound card 14
Figure 2.3. Demonstration of sampling mechanism: (a) in global shutter - each row is exposed simultaneously; (b) in rolling shutter - different rows in a frame are exposed at distinct time instances; $\Delta t1$, $\Delta t2$ and $\Delta t3$ represent respectively reset time, exposure time, and readout time per row 15
Figure 2.4. Drop in luminance samples at each exposed frame owing to idle period implementation: M denotes the illumination samples count that the sampling mechanism of rolling shutter can manage to expose during a frame time in the case that no idle period is present, and $M - L$ represents the samples count ignored when an idle period is implemented..... 17
Figure 2.5. Sampling mechanism of rolling shutter: (a) time domain illustration; (b) representation in a model of poly-phase decomposition..... 18
Figure 2.6. The operational blocks at the branch l of the model of poly-phase decomposition demonstrated in Figure 2.5 (b) 18
Figure 2.7. Sensor pattern noise components 22
Figure 3.1. Superpixels of an exemplary segmented image 28
Figure 3.2. The frequency spectrum obtained for an 30 fps video recorded in EU, i.e, 50 Hz nominal ENF: (a) when there is no idle period, the ENF appears at nominal illumination frequency, i.e., 100 Hz; (b) when idle period of 45% is implemented, new ENF components are derived in certain frequencies – the estimated ENF power also decreases noticeably owing to the idle period .. 33
Figure 3.3. Frequency shift of primary ENF depending on the idle period for videos of 30 fps, and for a mains power network of 50 Hz: (a) the developed analytical model; (b) simulation..... 34
Figure 4.1. ENF signals acquired from different points in Turkey's interconnected power network: (a) Bursa vs. Ankara; (b) Bursa vs. Konya; (c) Bursa vs. Kırıkkale 41
Figure 4.2. Household devices emitting acoustic mains hum: (a) boiler; (b) cooker hood; (c) vacuum cleaner 42
Figure 4.3. A comparison of acoustic mains hum interference into: (a) dynamic micro-phone; (b) electret microphone on spectrogram..... 44
Figure 4.4. Estimated ENF signals from an audio recorded: (a) by dynamic microphone; (b) electret microphone - the only ENF source in the recordings is the mains hum..... 45
Figure 4.5. Assessment of the proposed superpixel based ENF presence detecting method when representative ENF vector is obtained via the median operation: ROC curves were obtained for videos half of which were captured by CCD sensor, and the others by CMOS sensors 47
Figure 4.6. Assessment of the proposed superpixel based ENF presence detecting method when representative ENF vector is obtained via the mean operation: ROC curves were obtained for videos half of which were captured by CCD sensor, and the others by CMOS sensors 48

Figure 4.7. (a) Testing vertical phase method on a 480p video of wall-scene, video-1, resulting in an estimated idle period about 48%; (b) testing vertical phase method on a 720p video of wall-scene, video-2, resulting in an estimated idle period about 38% - both videos were captured in Turkey (50 Hz nominal ENF) at 30 fps by a Canon SX230HS model camcorder 50
Figure 4.8. (a) Testing vertical phase method on a 30 fps video of wall-scene, video- 3, resulting in an estimated idle period about 4%; (b) testing vertical phase method on a 23.976 fps video of wall-scene, video-4, resulting in an estimated idle period about 23% - both videos were captured by a Nikon D3100 model camcorder at 720p resolutions 51
Figure 4.9. Vertical phase method performed for a 23.976 fps-720p video of wall-scene recorded in Turkey by the Nikon D3100 under CFL bulb illumination, video 5: computation of the phase is a great challenge for this video 52
Figure 4.10. (a) Fourier Spectrum of the 30 fps-480p video captured by the Canon SX230HS, video-1; (b) constructed model of reference: variation in position of the pri- mary ENF harmonic depending on idle duration in the same video - the range of idle duration was obtained between 45% and 50% (The measured idle by the vertical phase approach was 48%) 53
Figure 4.11. (a) Fourier Spectrum obtained for the 30 fps-720p video captured by the Canon SX230HS, video-2; (b) constructed model of reference: variation in position of the primary ENF harmonic depending on idle duration in the same video - the range of idle duration was obtained between 35% and 40% (The measured idle by the vertical phase approach was 38%) 54
Figure 4.12. (a) Fourier Spectrum obtained for the 30 fps-720p video taken by the Nikon D3100, video-3; (b) constructed model of reference: variation in position of the primary ENF harmonic depending on idle duration in the same video - the range of idle duration was obtained between 0% and 5% (The measured idle by the vertical phase approach was 4%) 55
Figure 4.13. (a) Fourier Spectrum obtained for the 23.976 fps-720p video taken by the Nikon D3100, video-4; (b) constructed model of reference: variation in po- sition of the primary ENF harmonic depending on idle duration in the same video - the range of idle duration was obtained between 20% and 25% (The measured idle by the vertical phase approach was 23%) 56
Figure 4.14. (a) Fourier Spectrum obtained for the 23.976 fps-720p video taken under CFL light by the Nikon D3100, video-5; (b) constructed model of reference: vari- ation in position of the primary ENF harmonic depending on idle duration in the same video the range of idle duration was obtained between 20% and 25% (The measured idle by the vertical phase approach was 23%) 57
Figure 4.15. (a) Fourier Spectrum obtained for the 25 fps- 720p video taken by the Nikon D3100, video-6; (b) constructed model of reference: variation in position of the primary ENF harmonic depending on idle duration in the same video - the range of idle duration was obtained between 20% and 25% - alias ENF is detected at DC component for this frame rate, hence idle period for this case cannot be estimated via the vertical phase method..... 58
Figure 4.16. (a) Sample frames from an indoor video with moving content; (b) sample frames from an outdoor video with moving content 59

Figure 4.17. Spectra for different light sources 68
Figure 4.18. Computed ROC curves for all natural non-stabilized YouTube videos at 1080p in vision dataset - natural vs. natural 78
Figure 4.19. Computed ROC curves for all natural non-stabilized YouTube videos at 720p in vision dataset 79
Figure 4.20. Computed ROC curves for all natural, i.e., non-stabilized WhatsApp videos in Vision dataset - natural vs. natural 81

TABLES

Page

Table 4.1. The performance of the proposed superpixel based ENF presence detecting method when representative ENF vector is obtained via the median operation: computed AUC values for the ROC curves 47
Table 4.2. The performance of the proposed superpixel based ENF presence detecting method when representative ENF vector is obtained via the mean operation: computed AUC values for the ROC curves 48
Table 4.3. Tabulated results for the computed reference model for the frequency shift of the main ENF harmonic depending on idle duration for a video captured at 29.97 in EU 60
Table 4.4. Tabulated results for the computed reference model for the frequency shift of the main ENF harmonic depending on idle duration for a 25 fps video recorded in EU 61
Table 4.5. The settings and computed findings for each processed video for idle period estimation based on the proposed technique 63
Table 4.6. The statistics for the idle period estimation for 29.97 fps-720p videos captured by different cameras based on the proposed method..... 64
Table 4.7. Assessment of the proposed method for time-of-recording verification on native wall-scene videos..... 65
Table 4.8. Assessment of the proposed technique for time-of-recording verification on compressed wall-scene videos 66
Table 4.9. Assessment of the introduced method for time-of-recording verification on videos with moving content 67
Table 4.10. The light sources used in the experiments 69
Table 4.11. Assessment of time-of-recording detection for light sources of different types: computed AUC values..... 70
Table 4.12. True recording-time estimations rate (%) for light sources of different types 70
Table 4.13. Assessment of recording-time detection for compression in different levels and type: computed AUC values for LED 71
Table 4.14. True recording-time estimations rate (%) for compression in different levels and type for LED..... 71
Table 4.15. Assessment of recording-time detection for compression of different levels and type: computed AUC values for CFL..... 72
Table 4.16. True recording-time estimations rate (%) for compression in different levels and type for CFL 72
Table 4.17. Assessment of recording-time detection for LED for ground-truth ENF data of different lengths: computed AUC values..... 73
Table 4.18. True recording-time estimations rate (%) for LED for ground-truth ENF data of different lengths..... 73
Table 4.19. Assessment of recording-time detection for CFL for ground-truth ENF data of different lengths: computed AUC values..... 74
Table 4.20. True recording-time estimations rate (%) for CFL for ground-truth ENF data of different lengths..... 74

Table 4.21. The list of cameras and the set of non-stabilized YouTube videos exploited in the experiments 76
Table 4.22. The list of cameras and the set of non-stabilized WhatsApp videos exploited in the experiments 77
Table 4.23. AUC for All 1080p YouTube videos for different scenarios 78
Table 4.24. AUC for each single 1080p YouTube video for 2 different scenarios, natural vs. natural, and flat vs. natural 79
Table 4.25. AUC for all 720p YouTube videos for different scenarios 80
Table 4.26. AUC for each single 720p YouTube video for 2 different scenarios, natural vs. natural, and flat vs. natural 80
Table 4.27. AUC for all WhatsApp videos 81
Table 4.28. AUC for each single WhatsApp videos for 2 different scenarios, natural vs. natural, and flat vs. natural 82

1. INTRODUCTION

1.1. Motivation

In the information age, the number of digital media has grown rapidly as digital cameras have been anywhere. In particular, the boom in use of smartphones and of social media have led wide spread of digital image, audio and video. On the other hand, the digital age also led the digital recordings be straightforwardly edited and manipulated, or be artificially created by numerous techniques offered by various media editing tools. Moreover, the metadata which provides significant information about the recording such as media create date and time can be falsified by these tools. The manipulations can cause severe consequences for some circumstances, such as when the media is attempted to use as an evidence of a law court, or when they are used for defaming a well known person, i.e. a celebrity, a politician, etc. In view of these circumstances, media forensics has gained enormous significance for investigating and verifying the origin, integrity, authenticty and credibility of the media and the meta data. The development of forensic tools for different tasks, such as watermarking (Mettripun et al. 2013, Tachaphetpiboon et al. 2014), forgery detection (Naumovich and Memon 2003, Bayram et al. 2009), camera identification (Fridrich 2009, Sencar and Memon 2013), etc., has become a research field of great importance. The forensic tools based on ENF (Electric Network Frequency) and PRNU (Photo Response Non Uniformity) are among the most considerable and the evolving ones.

ENF is voltage frequency of mains electricity, and contrary to popular belief, it shows a continuous fluctuation in time depending on the gap between demanded and supplied power. ENF is intrinsically integrated into audio and video recordings in certain conditions and can be estimated from these recordings. The phenomenon that the ENF signal acquired from any power outlet in an interconnected network, at a particular time period, can be utilized as a reference signal, i.e., ground-truth, along with the fact that it can be extracted from digital media files has led to the exploitation of ENF in media forensics. ENF can be used for various forensic tasks for various scenarios, including time-of-recording verification of audio and video files. For instance, innocence of an alleged criminal can

be proved if a video of them can be found, and if the recording date and time of this video can be verified to overlap with those of the crime. It can also serve as getting date of time, originality and integrity of the videos of the past, which have been secret or unknown for some time for some reasons, emerge or come to light. This system can also be successfully used to test the audio-video integrity of a video by estimating the ENF from both the audio component and visual component. It is possible to increase the applications of ENF in media forensics.

PRNU is distinct response of each photo-sensor to light caused by non-homogeneous structure of silicon wafers and unavoidable defects in the manufacturing process. This unique characteristic of each camera sensor, even of those produced from the same wafer, and hence of the same brand and model, makes PRNU to be treated as a sensor fingerprint. PRNU noise is inherently embedded in each image or video frame recorded by a camera and can be estimated from them via some filtering operations. PRNU can be used mainly for identification of the source device by which a specified image or video is recorded. Since digital media is also an efficient means of committing criminal activities such as propaganda of terrorism, film piracy, child pornography, etc., identifying the source device correctly is significant. A true device attribution may consequently lead identifying the camera owner who may be the criminal, or the witness. It may be possible particularly by tracing the suspected social media accounts and detecting the locations where the sharings are made.

As can be realized from the above mentioned points, this dissertation mainly focuses on the two distinct topics: ENF based multimedia forensics, and PRNU based source camera attribution.

1.2. ENF (Electric Network Frequency) Forensics

ENF (Electric Network Frequency) instantly oscillates between the upper and lower limits of a nominal value (50/60 Hz) due to a constant imbalance between energy produced and energy consumed (Bollen and Gu 2006). In an interconnected power network, the load control mechanism used to stabilize the power grid prevents ENF (Electric Network

Frequency) from going beyond a certain limit around the nominal frequency. It also allows ENF intrinsically to show almost the same oscillation across the entire network (Bollen and Gu 2006). Accordingly, ENF signal measured at any power point on such a network can be utilized as reference, i.e. ground truth for the entire network (Grigoras 2005).

Grigoras (2005) discovered that ENF can be caught by an audio recorder in close proximity to mains-powered devices or to transmission cables due to electromagnetic wave propagation. He experimented and validated that ENF signal obtained directly from a mains power outlet is very similar to an accurately estimated audio ENF captured at the same time period. He proposed the ENF criterion which can be utilized to analyze digital audio recordings for a number of forensic applications including the integrity check, the recording date and time verification, and the recording region identification (Grigoras 2005, 2007, Garg et al. 2013). A similarity test between the estimated recording ENF and the ground-truth ENF plays a significant role in the ENF media forensics.

Brixen (2007) argued that ENF trace cannot be detected in an audio recording that is made using an electret microphone, equipped by numerous devices including smartphones, even in the existence of a very powerful electromagnetic field. Luckily, he suggested that it is possible to estimate ENF through an electret microphone recording made in the existence of acoustic mains hum, which is emitted by some devices, including household appliences. Chai et al. (2013), Fechner and Kirchner (2014), and Vatansever and Dirik (2017) experimented and validated this phenomenon by analyzing audio recordings made by some smartphone models using this acoustic noise produced by some household devices.

It has been found in the latest years that ENF is also present in a video captured under a mains-powered light source illumination (Garg et al. 2011). Such a light source's illumination intensity fluctuates depending on the ENF variation in the electricity network. The alterations in the luminance are inherently embedded in recorded videos. These alterations can be extracted from video by carefully analyzing subtle brightness changes along successive video frames, consequently resulting in extraction of ENF from video (Garg

et al. 2011, 2013, Su et al. 2014b, Wu et al. 2015, Hajj-Ahmad et al. 2016, Vatansever et al. 2017, Vatansever et al. 2019a).

Most CMOS (Complementary Metal Oxide Semiconductor) equipped consumer cameras use sampling system of a rolling shutter mechanism to expose a video frame while those with CCD (Charged Coupled Device) sensors exploits sampling system of a global shutter mechanism. In the global shutter, frame rows are exposed simultaneously, while each line is caught at separate times with a rolling shutter. This distinction in sampling processes has resulted researchers to build distinct techniques in extracting ENF in videos. The first technique (Garg et al. 2011) is founded on the phenomenon of global shutter. It simply uses an average of all the stable pixels in a frame. Accordingly one illumination sample is obtained per frame, leading to a sampling rate as many as *video frame rate*. Hence, it is incapable of meeting Nyquist theorem (Proakis and Manolakis 2007). It is therefore reliant on the alias. The other method is intended for the rolling shutter system. It is on the basis of averaging stable pixels in each row, consequently leading to one sample of illumination per row (Su et al. 2014b, Wu et al. 2015). This technique utilizes the benefit of the enhanced frequency of sampling in the rolling shutter, which is the *number of rows* $\times$ *video frame rate*. The rolling shutter strategy, however, carries with it the problem of idle period between two consecutive frames where sampling is not performed. Therefore, some samples of illumination at the end of each frame are lost. Su et al. contributed to a basic concept ant comprehension on how the primary ENF is shifted to certain frequencies owing to the idle time. (Su et al. 2014b). For their analysis, they used a filter bank model.

1.3. PRNU (Photo Response Non-Uniformity) Forensics

The inhomogenity of the semiconductor material (silicon wafer) from which a camera sensor is manufactured, and the imperfections in the production process cause the light sensitivity of each cell of photo-sensor be different from each other. This unique characteristic of camera sensors is called PRNU (Photo Response Non Uniformity), and it can be regarded as a sensor fingerprint since PRNU of even the same make and model cameras are different (Lukáš et al. 2006, Chen et al. 2007, Filler et al. 2008). PRNU noise is intrinsically embedded into every recorded image, and it can be estimated from these

images by means of a set of filtering operations. To identify or verify the source device of a test image, a similarity test between reference PRNU pattern of a camera and PRNU noise estimate of the given image is required. Reference PRNU pattern of a camera can be computed by roughly averaging the PRNU noises obtained from a set of images images recorded by the same device. The more images lead to the better reference PRNU estimate.

The fact that digital video files consist of a large number of frames may give rise to the idea at first glance that the source device identification for video is much easier and reliable than that for image. However, the encoding of digital video frames in high compression rates refutes this thinking as it contaminates the PRNU noise to be estimated. The challenge is doubled if motion compensation is also applied to videos. That is, for PRNU-based source identification to work correctly in digital videos, it is necessary to know with which cells of photo-sensor and the pixels in the video frames match. However, motion compensation algorithms disrupt the synchronization of video frames and cause pixel shifts in location between consecutive frames. Consequently, PRNU-based source camera identification task is actually more challenging for videos than that for images.

The pioneering work in the field of PRNU based video forensics is that of Chen et al. (2007) who studied whether two video were taken with the same camera. In this study, they targeted on removing the quantization noise, i.e., blockiness artifacts on the estimated PRNU to improve the performance. However, for poor quality videos, i.e. in low bit rates, they required very long videos, i.e. longer than 10 minutes for the algorithm to work efficiently. Another pioneering work is the research conducted by Mondaini et al. (2007) for detecting forgery in a given video. They assumed there can be no modification in the initial frames of a video. Hence, PRNU reference pattern was estimated by taking the first 50 frames of the test video, and it was improved with the PRNU noises of other frames depending on a similarity test between this pattern and those of other frames. However, their assumption is not always the case, i.e., it is likely that tampering can also be in the initial frames.

Chuang et al. (2011) investigated how the frame type in compressed videos affects the PRNU-based source identification task. Their experiments showed I-frames are more dependable in PRNU estimation than P-frames . However, in the same study, it was also resulted that using only I-frames from the test video in PRNU noise estimation is insufficient. Hyun et al. (2012) emphasized that using the MACE (Minimum Average Correlation Energy) filter is more robust against noise rather than NCC (Normalized Cross Correlation), and it significantly improves the detection rate of the source camera when calculating the similarity between PRNU patterns of highly compressed and low resolution videos. Al-Athamneh et al. (2016) claimed that using only the G channel significantly increases the success rate, rather than using all the R, G and B channels in the PRNU estimation. However, to the best of our knowledge, no supporting argument was provided by any researcher after their work.

Taspinar et al. (2016) worked on out-of-camera stabilized videos (FFMPEG deshaker, etc.). They aimed to align the stabilized video frames for an accurate PRNU noise estimation. In order to obtain the transformation parameters used in the stabilization process, they implemented inverse affine transformation based on the PRNU noise obtained from the first frame.

As the images and videos captured by the same source device have different resolutions, the photo-receiving cells of the video frames and image pixels will not directly match each other. Therefore, the PRNU trace obtained from the images should be made comparable with the PRNU noise of the test video with a particular geometric transformation approach. Shullani et al. (2017) and Iuliani et al. (2017) were influenced by the work of Goljan and Fridrich (2008), who proposed a method for the determination of conversion parameters within the scope of recognizing the source of the cut and scaled images. Shullani et al. (2017) and Iuliani et al. (2017) targeted on the task whether the test images and videos can be taken with the same camera.

Kouokam and Dirik (2019) proposed a block based PRNU noise estimation approach for non-stabilized videos. By benefiting from inverse DCT (Discrete Cosine Transform) transform, they compute a frame mask for each frame to label the pixels as degraded

PRNU, or useful PRNU. Although, the proposed approach do not contribute to the performance for the native videos, it considerably increases the performance for YouTube videos. The PRNU related part of this dissertation is mainly based on their suggested technique.

1.4. Main Contributions and Dissertation Organization

In this thesis, we provide a comprehensive study on ENF (Electric Network Frequency) based audio and video forensics and present a number of novel methods (Vatansever and Dirik 2016, Vatansever and Dirik 2017, Vatansever et al. 2017, Vatansever et al. 2019a,b). We also provide a comparative analysis on PRNU based source camera attribution for social media video pairs.

First of all, we present the research on the sources of ENF, i.e., electromagnetic field and mains hum, in audio (Vatansever and Dirik 2016). We investigate how ENF is integrated into audio files or audio content of video files depending on the microphone type, i.e., dynamic or electret. Moreover, it is investigated if any trace of information can be obtained related to the recording scene and/or recording media by analyzing the estimated ENF signal from the audio recordings.

Secondly, we explore if ENF presence in video can be detected (Vatansever et al. 2017). It is essential to assess in ENF-based video forensics whether a video includes an ENF trace before proceeding with further assessment. For example, if there is no trace of ENF in a given video, or if it contains a poor quality ENF signal,i.e. low SNR value, searching the estimated signal in an ENF database for verification of the recording time or of region-of-recording would be useless. Indeed, if the lack of an ENF signal or a poor quality ENF can be detected by a rapid test, a significant amount of time as well as computational load is saved. To our knowledge, none of the other works provides a strategy that can detect existence of the ENF trace in a given video. In this thesis, a superpixel-based ENF detection method is presented. The proposed method performs ENF signal estimations from superpixels. Our motive to exploit superpixels for ENF estimation is that every pixel is almost uniform in brightness, texture and color in a superpixel area and therefore has

uniform characteristic of reflectance (Achanta et al. 2010). Hence, use of such a region makes it possible to estimate ENF signal from videos exposed by both CCD and CMOS sensor. A "so-called ENF signal" is extracted separately from each stable superpixel. The reason to use the word "so-called ENF" is that it is not initially known that the estimated signal is really an ENF signal. Depending on the resemblance of the estimated "so-called ENF" signals, it can be concluded that whether or not an ENF trace is contained by the given video. It is notable that the proposed technique does not require any reference ENF database, i.e. ground truth.

Third, the phenomenon presented in (Su et al. 2014b) is researched further and an analytical model is developed to investigate how mains-powered illumination frequency, hence the primary ENF, is attenuated, and shifted in relation to the duration of the idle period (Vatansever et al. 2019a). Among the emerging components of the ENF resulted by a specific idle period,the greatest two are investigated.

Next, a new method for idle period estimation is suggested (Vatansever et al. 2019a) based on the proposed aforementioned model. To our knowledge, the work of Hajj-Ahmad et al. (2016) is the only research using ENF in camera forensics except for ours. They depend primarily on calculating the time required to read a single frame. They perform ENF phase estimation for each row, i.e., vertical phase analysis (Hajj-Ahmad et al. 2015a). They treat the sequence of the mean intensity for an i^{th} row of all consecutive frames as a distinct time series and compute the phase of ENF for this row, resulting in in a sampling frequency as the same as *video frame rate*. Accordingly, it does not meet the Nyquist criteria (Proakis and Manolakis 2007), therefore relies on alias ENF. However, alias ENF is obtained at 0 Hz for the video frame rates being a nominal ENF divisor. Hence, their method cannot work on such videos, e.g. 25 fps videos for 50 Hz network (EU) or 30 fps videos at 60 Hz network (US). Operating in a noisy video, where the captured ENF's power is weak for computation of phase accurately, is also a challenge to their method. Our suggested method for idle-period computation can handle these weaknesses.

Our developed model also contributes to a new time-of-recording verification approach (Vatansever et al. 2019a) that works better than the other methods (Garg et al. 2013, Su

et al. 2014b), (Wu et al. 2015). A careful search for emerging components of ENF resulted from the idle period, followed by presumptions of idle period for each emerged component and interpolation for each presumption lead to more accurate ENF signal estimates. A more accurate ENF signal accordingly results in higher performance in time-of-recording verification.

Another research provided in this thesis is the study on factors affecting ENF forensics for video (Vatansever and Dirik 2017, Vatansever et al. 2019b). Data characteristics and ambient conditions in the recording environment can have a significant effect on the quality of ENF in video. These circumstances may therefore affect the efficiency of ENF based forensic tasks such as time-of-recording verification (Garg et al. 2013, Vatansever et al. 2019a), multimedia synchronization (Su et al. 2014a,c), camera characterization (Hajj-Ahmad et al. 2016, Vatansever et al. 2019a), and ENF presence detection (Vatansever et al. 2017). It is mainly investigated that how distinct sources of illumination, distinct compression ratios, and reference ENF signal (ground-truth) of different lengths affect the efficiency of the video time-of-recording estimation.

In this thesis, we also conduct a study on PRNU (Photo Response Non Uniformity) based source camera attribution for social media videos. For this purpose, the proposed technique in (Kouokam and Dirik 2019) is further researched and a comparative and complementary analysis is provided via the experimental work on YouTube and WhatsApp videos.

To sum up, this thesis's primary contributions include: (1) a comparative study between sources of ENF in audio and the link with the microphone type; (2) proposal of a superpixel based ENF presence detection algorithm; (3) creation of a model for where the frequency of the primary video ENF harmonic is shifted and how the captured ENF's power is attenuated owing to the idle period in rolling shutter mechanism; (4) development of a new method for idle period extraction; (5) exploration of a new technique for time-of-recording verification task for rolling shutter exposed videos; (6) an analysis on main factors affecting ENF based video forensics; (7) a comparative analysis on PRNU based source camera attribution from social media video pairs.

The rest of this thesis is organized as follows: In Section 2, we provide the foundations and the main concepts related to ENF and PRNU based media forensics. In Section 3, we present the methods we developed. In Section 4, we provide the experimental findings. Finally, the conclusion is given in Section 5.

2. THEORETICAL BASICS AND LITERATURE REVIEW

In this section, theoretical background, concepts and foundations related to the thesis subject are provided. As the thesis covers two distinct topics, namely ENF based multimedia forensics and PRNU based source device identification, all the relevant concepts are presented under subsections of the two main topics.

2.1. ENF based Media Forensics

The property that ENF signal measured at any point across an interconnected electricity network can be utilized as reference, together with the fact ENF signal can be estimated from digital audio and video have led to the exploitation of ENF in media forensic. The ENF criteria can be utilized for various forensic and anti-forensic applications such as time-of-recording verification (Fechner and Kirchner 2014, Garg et al. 2013, Bykhovsky and Cohen 2013), media authentication (Hua et al. 2016, Savari et al. 2016), multimedia synchronisation (Su et al. 2014a), (Su et al. 2014c), (Chuang et al. 2013), power grid identification (Hajj-Ahmad et al. 2015b) and camera read-out time estimation (Hajj-Ahmad et al. 2016).

2.1.1. ENF power model

Instantaneous voltage of a mains electricity can be expressed as:

$$\begin{aligned} V(t) &= \sqrt{2}V_0\cos(\phi(t)) \\ &= \sqrt{2}V_0\cos(2\pi f_n t + \theta(t) + \alpha) \\ &= \sqrt{2}V_0\cos(2\pi f_n t + 2\pi\int_0^t f_e(\tau)\mathrm{d}\tau + \alpha) \end{aligned} \tag{2.1}$$

where V_0 represents effective mains voltage, f_n is nominal ENF, and α denotes initial phase offset (Bollen and Gu 2006). $f_e(t)$ denotes instantaneous alteration from nominal ENF. $\theta(t)$ is the instant phase varying depending on the supply and demand imbalance. From Equation 2.1, the instantaneous frequency at a t moment can be written as:

$$f(t) = \frac{1}{2\pi}\frac{d\phi(t)}{dt} = f_n + f_e(t) \tag{2.2}$$

As f_n is constant, ENF varies based on $f_e(t)$ alterations. $f_e(t)$ can be written as (Bollen and Gu 2006):

$$f_e(t) = \frac{f_n}{2H}(P_s(t) - P_d(t)) \tag{2.3}$$

where $P_d(t)$ represents demanded power, $P_s(t)$ is supplied power, H denotes an inertia constant. Accordingly, $f_e(t)$ and $f(t)$ vary depending on the instant disparity between the consumed power and the generated power.

2.1.2. ENF in audio

One of the most remarkable tools developed over the last decades for forensic analysis of digital audio recordings is the ENF criterion proposed by Grigoras (Grigoras 2005). Electric frequency alterations of a mains-powered network can be captured by an audio recording made by a dynamic microphone, i.e. equipped with a moving coil - magnetic microphone, in existence of substantial amount of electromagnetic field sourced by the mains power voltage (Grigoras 2007). However, if an electret microphone, which contains capacitive material instead of coil-magnet pair, is used in the same setting instead of a dynamic one, no ENF trace is found in the recorded audio (Brixen 2007). Luckily, ENF-induced electromagnetic propagation can lead to an acoustic network noise, i.e. mains hum, to be generated and emitted by some devices caused by vibrations in the electronic circuit elements, such as transformer, in these devices. Such a noise can be captured by audio recordings, and consequently it becomes possible to estimate ENF from the audio recordings made by electret microphone in the presence of acoustic mains (Chai et al. 2013). As mobile phones are equipped with electret microphones, it is essential for an audio recording to be made by such a device in the existence of the mains hum so that ENF can be captured.

If ENF signal in an audio file can be estimated accurately, it shows a great similarity with the ground truth ENF. An analysis of the the extent of this similarity, generally by using normalized cross-correlation (NCC), may provide noteworthy forensic evidence about the audio, and/or the recording region such as that it can be determined or verified in which

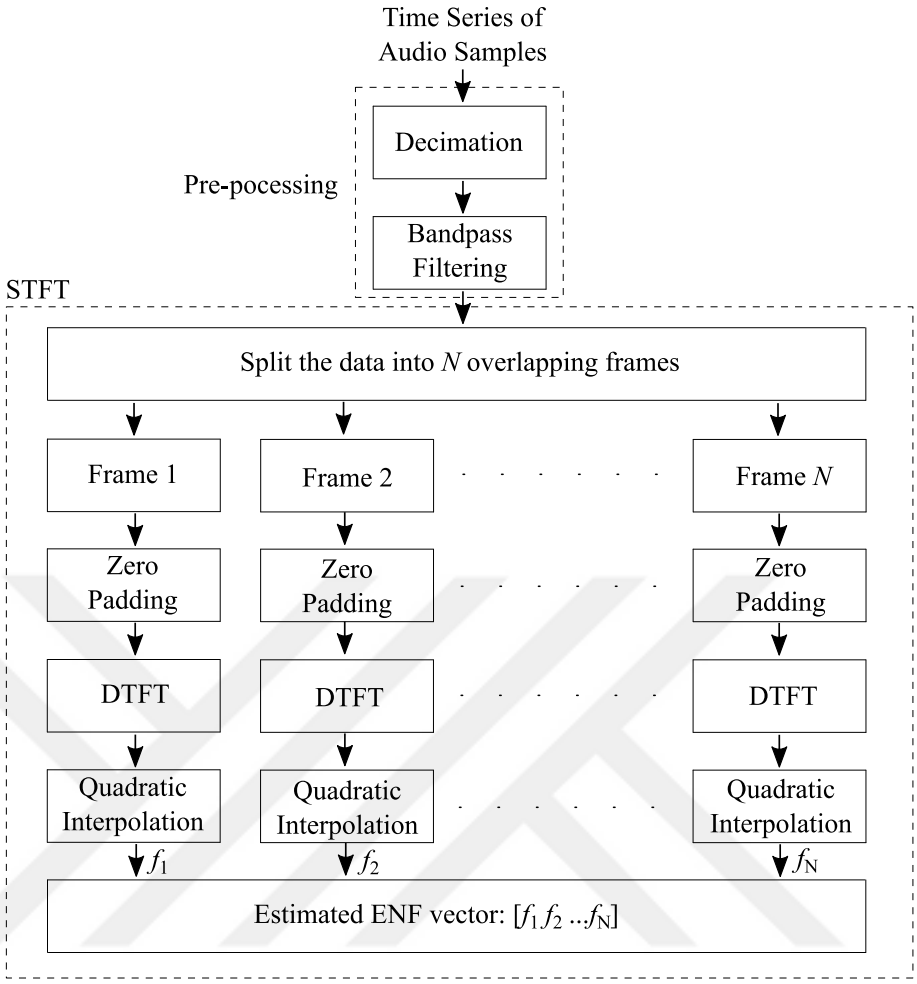


Figure 2.1. Overall ENF extraction procedure from audio

region, on which day and time the test recording is made (Grigoras 2005, Garg et al. 2013). That is, the extracted ENF vector can be searched in a ground-truth ENF database, i.e., reference ENF, and depending on the value of the peak correlation coefficient, and on the corresponding lag point, recording region and recording day-and-time of a test video can be detected or verified (Grigoras 2005). The ground truth database of ENF can be acquired from any power outlet in the grid network, which is discussed in Section 2.1.4.

2.1.3. ENF estimation for audio

ENF can be estimated from audio by using any time or frequency domain methods discussed in (Grigoras 2005, Cooper 2008, Grigoras et al. 2009, Hajj-Ahmad et al. 2012). Due to the time and performance efficiency, Short Time Fourier Transform (STFT) based

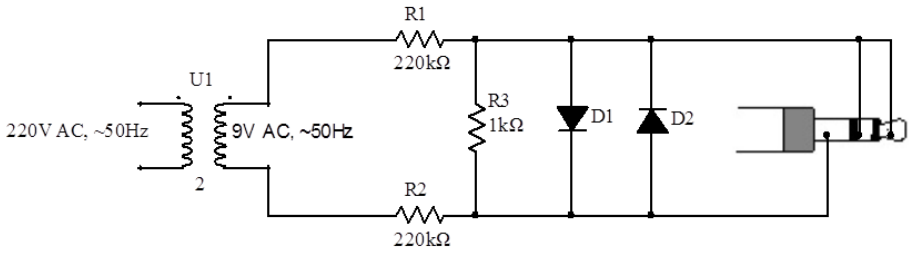


Figure 2.2. The intermediate circuit for extraction of ground-truth ENF directly from mains power outlet via sound card

frequency estimation technique is generally preferred in ENF estimation from audio. However, before implementation of STFT on the audio, it is pre-processed first for efficiency. For this purpose, the acquired audio samples are decimated first into 200 Hz, which satisfies Nyquist criteria (Proakis and Manolakis 2007) for 50 Hz nominal frequency. A band-pass filter around the ENF region of interest is then applied, preferably for the range between 49 Hz and 51 Hz. From this point on, STFT can be computed for the pre-processed audio samples by using 20 seconds' data frames with one second shift, i.e., 19 seconds overlaps. Quadratic interpolation (Cooper 2008) of the peak value for each of the DTFT applied frame is then performed, consequently leading 1-second/sample ENF resolution. Overall ENF estimation procedure for audio is depicted in Figure 2.1.

2.1.4. Ground-truth ENF acquisition

Ground truth ENF database, i.e., reference ENF, is required particularly for time-of-recording verification task. Hence, it is essential to instantaneously acquire the electric network frequency from a power outlet and save it. For this purpose, an ENF adapter circuit such as that shown in Figure 2.2 can be used. One end of such an adaptor is connected to a power outlet, and the other end is connected to a PC sound card. Accordingly, the reduced mains voltage (20mV) signal can be transferred to the PC sound card via the TRS (microphone) type connector. This signal is then pre-processed and is applied the STFT (Short Time Fourier Transform) based frequency estimation algorithm as in Section 2.1.3. Accordingly, instantaneous changes of the network frequency is computed and recorded in real time.

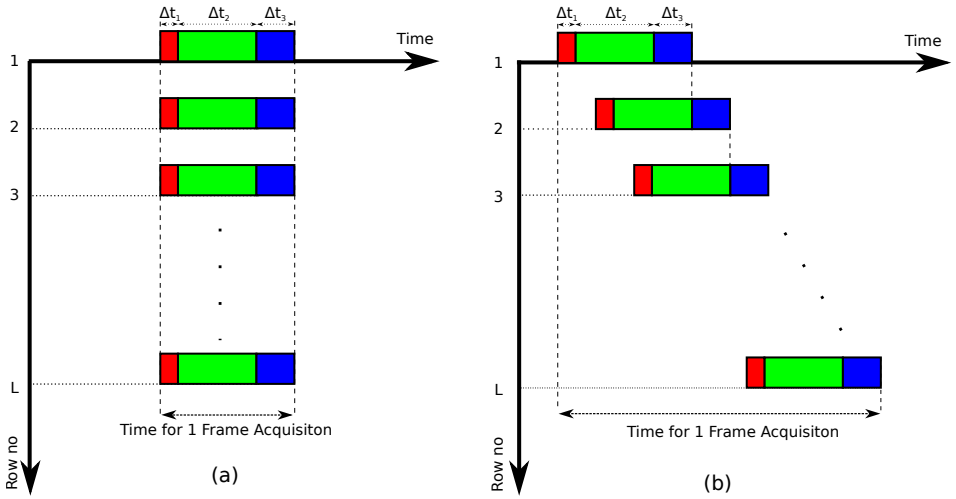


Figure 2.3. Demonstration of sampling mechanism: (a) in global shutter - each row is exposed simultaneously; (b) in rolling shutter - different rows in a frame are exposed at distinct time instances; Δt_1, Δt_2 and Δt_3 represent respectively reset time, exposure time, and readout time per row

2.1.5. Light source flicker and ENF in video

Luminance intensity of a mains-powered source of light fluctuates depending on ENF alterations. Since light sources flicker at both negative and positive cycle of mains power voltage, the frequency of illumination is twice the ENF. The luminance signal can therefore be regarded as the absolute of the cosine function in Equation 2.1. In Europe, for instance, nominal ENF is 50 Hz, so the illumination frequency alternates around 100 Hz.

2.1.6. ENF estimation for video

Short Time Fourier Transform (STFT) is also preferred for estimation of ENF from the video recordings. However, in videos, the input data for the STFT is the time-series for variation of luminance samples rather than the audio in Figure 2.1. Moreover, this time-series cannot be directly obtained as audio samples, and it highly depends on the camera sampling mechanism. Most consumer cameras are equipped with either a CCD sensor, or a CMOS sensor. Although, a CCD sensor commonly uses sampling system of global shutter, a CMOS sensor mainly exploits sampling system of rolling shutter. In the global shutter, a whole frame is sampled at once, i.e., all rows in a frame is concurrently exposed, whilst in the rolling shutter, different rows of a single frame is captured consecutively at

separate instances of time. Figure 2.3 depicts the timing for exposure of a single frame for the two distinct sampling mechanisms (Gu et al. 2010). Not surprisingly, these difference in sampling mechanisms results in the exposure of the luminance data by cameras differently. Consequently, the ENF extraction techniques from video in the literature are mainly designed by considering these sampling processes, specifically in the stage of construction of the variation of illumination time-series. In the following subsections, the two distinct approaches for estimation of these time series are addressed.

Global shutter based approach

The proposed approach in (Garg et al. 2011) is based primarily on the global shutter phenomenon. As all pixels belonging to a frame is captured at once in the global shutter, one illumination sample per frame can only be acquired in this technique, by averaging all stable pixels per frame. The sampling rate for this method is therefore equal to the *video frame rate*. In order to be able to estimate an luminance frequency around 100 Hz (double the ENF, i.e. 50 Hz), at least 200 Hz sampling rate is normally required for reconstruction of a signal based on the criteria of Nyquist Sampling Theorem (Proakis and Manolakis 2007). However, in this approach, the sampling rate is most likely to be much lower than that as most consumer cameras do not offer such a high frame rate. Hence, this approach is reliant on operating with alias ENF. The aliased frequency of illumination, f_a can be obtained as follows (Hajj-Ahmad et al. 2015a):

$$f_a = |f_l - k \cdot f_s| < \frac{f_s}{2}, \quad \exists k \in \mathbb{N} \tag{2.4}$$

where f_s is the camera sampling frequency, i.e. video frame rate, and f_l denotes the nominal illumination frequency. Accordingly, when a 100 Hz illumination signal is sampled with 29.97 fps frame rate, for instance, the primary frequency of alias ENF is acquired at 10.09 Hz. It should be highlighted that although the frequency of illumination can still be estimated as alias, the alias frequency is obtained at 0 Hz for a video frame rate at a nominal ENF divisor. Therefore, use of this scheme under such a condition is a great challenge.

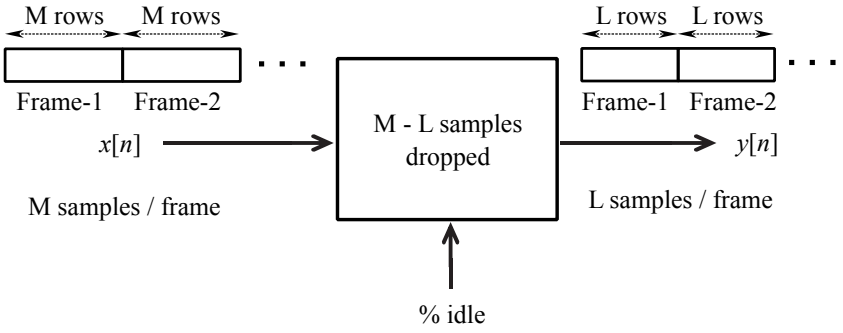


Figure 2.4. Drop in luminance samples at each exposed frame owing to idle period implementation: M denotes the illumination samples count that the sampling mechanism of rolling shutter can manage to expose during a frame time in the case no idle period is present, and $M - L$ is the samples count ignored when an idle period is implemented

Rolling shutter based approach

The technique in (Su et al. 2014b) and in (Wu et al. 2015) is based on the phenomenon of the rolling shutter. Since each raw belonging to a frame is recorded at sequential time instances in rolling shutter, a separate illumination sample for each row of a frame can be obtained in this technique, by computing the luminance shift from the mean value of each pixel intensity variation along the video, and averaging them (Wu et al. 2015). For ease of understanding, we simply call the average luminance shift for a row as row illumination sample. The time-series for variation of illumination along the video is constructed by concatenating the illumination sample of each row of all successive frames. Consequently, the sampling rate for this scheme is obtained as *frame rate* $\times$ *number of rows*. Such a sampling frequency is much greater than that Nyquist criteria requires. Alias ENF accordingly is not a case here.

2.1.7. Idle period effect on ENF

While the rolling shutter may result in an increase in the sampling rate, hence avoid alias frequency, it has the problem of idle period at each frame. That is, some luminance samples are missed during the idle time (Su et al. 2014b). A representation of this phenomenon is shown in Figure 2.4. Here, M denotes the rows count that can possibly be captured by the rolling shutter sampling mechanism of a camera during a frame time provided that no idle period is present, i.e no sample is missed. That is, this shutter system

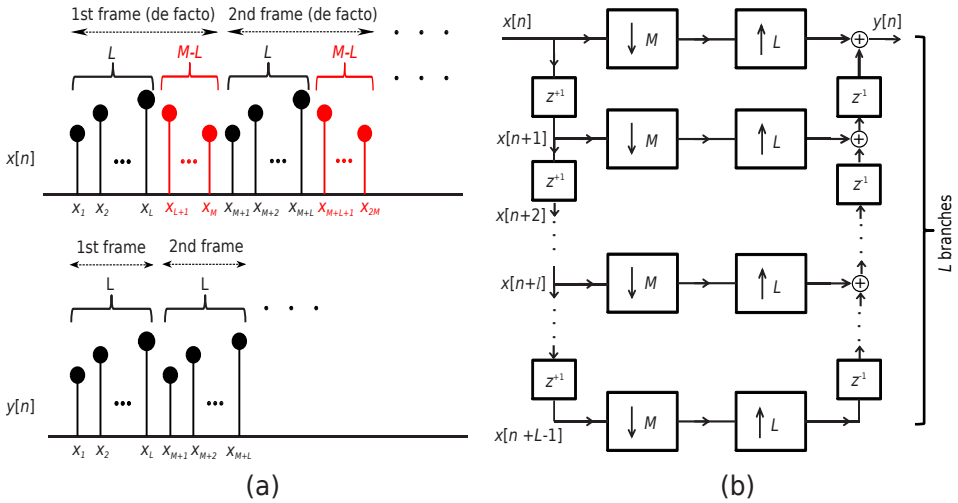


Figure 2.5. Sampling mechanism of rolling shutter: (a) time domain illustration; (b) representation in a model of poly-phase decomposition

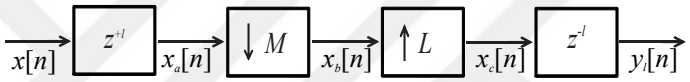


Figure 2.6. The operational blocks at the branch l of the model of poly-phase decomposition demonstrated in Figure 2.5 (b)

has a capacity of exposing M rows during a frame time in the case that no idle period exists. $M-L$ refers to the samples count ignored over the idle time. Accordingly, L denotes the remaining samples count, i.e. the rows count per frame that are actually exposed ($L \leq M$). $x[n]$ refers to the time series of illumination provided that no idle time exists. $y[n]$ represents the remaining time series of illumination after the sample losses in each frame's idle period.

A demonstration of the above operations in time domain is provided in Figure 2.5 (a) (Su et al. 2014b) and a polyphase decomposition of this representation is depicted in Figure 2.5 (b) (Su et al. 2014b). The anti-aliasing filter is omitted in this model for simplicity. The input signal is shifted back in time in each branch of the model, Equation 2.5, followed by an M-fold down-sampling filter, Equation 2.6. Then, an L-fold up-sampling filter is implemented, Equation 2.7. Next, in the appropriate order, the signal is shifted forward, Equation 2.8. These stages are illustrated for the lth branch in Figure 2.6. Discrete Time Fourier Transform (DTFT) of $x_a[n]$, $x_b[n]$, $x_c[n]$ and $y_l[n]$ are obtained accordingly

as follows:

$$X_a\left(e^{j\omega}\right) = X\left(e^{j\omega}\right)e^{j\omega l} \tag{2.5}$$

$$\begin{aligned} X_b\left(e^{j\omega}\right) &= \frac{1}{M}\sum_{m=0}^{M-1} X_a\left(e^{j\left(\frac{\omega}{M}-\frac{2\pi m}{M}\right)}\right) \\ &= \frac{1}{M}\sum_{m=0}^{M-1}\left[X\left(e^{j\left(\frac{\omega}{M}-\frac{2\pi m}{M}\right)}\right)e^{j\left(\frac{\omega}{M}-\frac{2\pi m}{M}\right)l}\right] \end{aligned} \tag{2.6}$$

$$\begin{aligned} X_c\left(e^{j\omega}\right) &= X_b\left(e^{j\omega L}\right) \\ &= \frac{1}{M}\sum_{m=0}^{M-1}\left[X\left(e^{j\left(\frac{\omega L}{M}-\frac{2\pi m}{M}\right)}\right)e^{j\left(\frac{\omega L}{M}-\frac{2\pi m}{M}\right)l}\right] \end{aligned} \tag{2.7}$$

$$Y_l\left(e^{j\omega}\right) = X_c\left(e^{j\omega}\right)e^{-j\omega l} \tag{2.8}$$

Then, discrete Fourier transform of the resulted signal at the end of l^{th} branch, $Y_l\left(e^{j\omega}\right)$ can be written as follows:

$$Y_l(\omega) = \frac{1}{M}\left(\sum_{m=0}^{M-1} X\left(\frac{\omega L - 2\pi m}{M}\right)e^{j\frac{\omega L-2\pi m}{M}l}\right)e^{-j\omega l} \tag{2.9}$$

In Equation 2.9, $Y_l\left(e^{j\omega}\right)$ is denoted in the form $Y_l(\omega)$ for simplicity. $Y\left(e^{j\omega}\right)$, $X\left(e^{j\omega}\right)$ and other related frequency domain variables are represented in similar format from this point on. Consequently, the output signal $Y(\omega)$ can be obtained by combining the signals formed at the end of each branch as follows (Su et al. 2014b):

$$\begin{aligned} Y(\omega) &= \sum_{l=0}^{L-1} Y_l(\omega) \\ &= \sum_{l=0}^{L-1}\frac{1}{M}\left(\sum_{m=0}^{M-1} X\left(\frac{\omega L - 2\pi m}{M}\right)e^{j\frac{\omega L-2\pi m}{M}l}\right)e^{-j\omega l} \\ &= \sum_{m=0}^{M-1} X\left(\frac{\omega L - 2\pi m}{M}\right)F_m(\omega) \end{aligned} \tag{2.10}$$

where

$$F_m(\omega) = \frac{1}{M} \sum_{l=0}^{L-1} e^{-j\frac{\omega(M-L)+2\pi m}{M} l} \tag{2.11}$$

2.1.8. Idle period estimation

Camera forensics is a research field of great importance for various applications including common source camera attribution of image/video pairs, and source camera identification of a suspicious image/video (Sencar and Memon 2013), (Shullani et al. 2017). The pioneering research using ENF for camera forensics is that of Hajj-Ahmad et al. (2016). Their technique is targeted to cameras equipped with rolling shutter system. Basically, it exploits the property of the shift in phase of ENF between successive rows owing to the distinct sampling of lines in rolling shutter. They essentially calculate the elapsed time for reading one frame, T_{ro} by computing the ENF phase per row, i.e. vertical phase as:

$$T_{ro} = \frac{L\tilde{\omega}_b}{2\pi\tilde{f}_e} \tag{2.12}$$

where L denotes the rows count per frame, $\tilde{f}_e$ is the fluctuating ENF component, and $\tilde{\omega}_b$ represents vertical radial frequency. $\tilde{\omega}_b$ can be estimated via computation of the slope of the line of vertical phase which can be formed by estimating the ENF phase, $\Phi[l]$ ($l \in \{1,2,3,...,L\}$, per row (Hajj-Ahmad et al. 2016). $\Phi[l]$ can be obtained by performing DTFT for the time-series of l th row which can be constructed through mean intensity of the l th row of consecutive frames. As the sampling rate for this time-series is as many as video frame rate, Nyquist criteria is not satisfied. Accordingly, peak search is performed around alias ENF in the frequency domain. Accordingly, the advantage of the increased sampling rate offered by the rolling shutter system, i.e., as many as *camera frame rate* $\times$ *number of rows*, addressed in Section 2.1.6, is lost in this approach, although it is tailored to videos exposed by this shutter system.

2.2. PRNU based Media Forensics

The phenomenon that PRNU is the unique characteristic in a camera, hence it is regarded as the camera fingerprint, along with the fact that PRNU noise estimate can be obtained from image and video have boomed the use of PRNU in media forensics, particularly in source camera identification. In this subsection, the related foundations and main concepts are provided.

2.2.1. Pattern noise of imaging sensors

There are various imperfections and sources of noise interfering into a camera's image acquisition system. Sensor pattern noise (SPN) remains essentially the same when several images of the exact same scene are recorded. SPN is embedded in every image the sensor takes. The main components of the SPN are photo response non-uniformity (PRNU), and fixed pattern noise (FPN) as shown in Figure 2.7. The FPN is formed due to dark currents and mainly refers to the difference between pixels when the sensor is not exposed to light (Lukáš et al. 2006). Some consumer cameras may automatically suppress the SPN noise via subtraction of a dark frame from each single image captured. The FPN may depend also on temperature and exposure. The PRNU is the dominant part of the SPN in natural images. It is mainly associated with pixel non-uniformity (PNU), which can be defined as distinct pixel sensitivity to light formed by the inhomogeneous nature of silicon wafers and the emerging imperfections, and defects in the sensor production process. The PNU noise's intrinsic nature makes it impossible that sensors manufactured from even an identical wafer will have associated PNU patterns (Lukáš et al. 2006). Accordingly, cameras of even the same make and model have different PRNU patterns.

An estimate of PRNU can be obtained from an image or a video via a set of filtering operations discussed in Section 2.2.2. In this thesis, the estimated PRNU pattern from image or video is referred to as PRNU noise. Roughly averaging the PRNU noises obtained from multiple still images or multiple video frames provide better estimate of PRNU. A good quality of PRNU noise estimate can be used as reference PRNU pattern in source camera attribution for a test image or video.

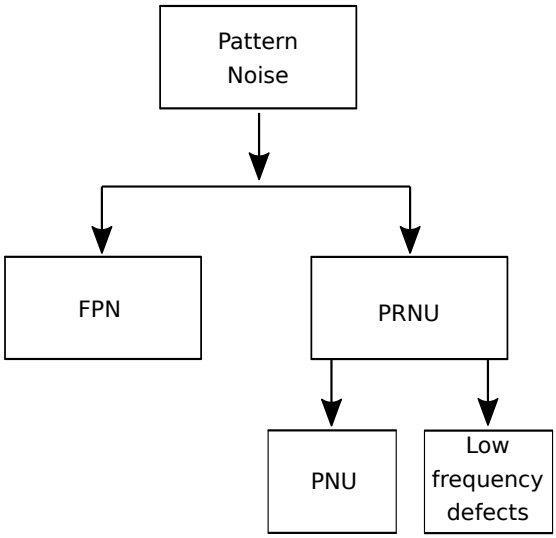


Figure 2.7. Sensor pattern noise components

2.2.2. PRNU estimation for image

Discarding the factor of gamma correction, the output image $\mathbf{I}$ produced by a digital camera can be expressed as follows (Dirik and Karaküçük 2014):

$$\mathbf{I} = \mathbf{I}_0 + (\mathbf{I}_0\mathbf{F} + \Theta) \tag{2.13}$$

where $\mathbf{F}$ is the PRNU fingerprint of a given camera; $\mathbf{I}_0$ is the true optical view, i.e., original image without any added noise component; Θ represents the sum of all other noise components other than PRNU. The PRNU noise estimate of an image $\mathbf{I}$ can be computed with the use of a wavelet-based noise filtering algorithm (WDF) as:

$$\mathbf{N} = \mathbf{I} - WDF\,(\mathbf{I}) \tag{2.14}$$

where $\mathbf{N}$ is the PRNU noise estimate for $\mathbf{I}$. By using a maximum likelihood estimator, an estimate of $\mathbf{F}$ can be obtained from the PRNU noise estimates of L images as follows:

$$\hat{\mathbf{F}} = \frac{\sum_{i=1}^{L} \mathbf{N}^{(i)}\mathbf{I}^{(i)}}{\sum_{i=1}^{L} (\mathbf{I}^{(i)})^2} \tag{2.15}$$

However, the estimated $\hat{\mathbf{F}}$ in Equation 2.15 contains some periodic and non-unique artifacts being present in every image. Color interpolation applied by CFA (Color Filter

Array), row-wise as well as column-wise processing in sensor, and blockiness artifacts caused by JPEG compression are the main sources of these residues (Chen et al. 2008). In order to remove these high similarity residues from the PRNU fingerprint estimate, and accordingly to minimize the similarity between the estimated PRNU noise of different cameras, the column average, for each column, is subtracted from every pixel in the same column. This is performed for each color channel individually. Similarly, the row average, for each row, is subtracted from every pixel in the same row. As a result of these row-column operations, which can be represented as $ZM(\hat{\mathbf{F}})$, the linear pattern is removed, and each row and column have zero mean (Chen et al. 2008). Finally, to remove visually identifiable patterns in $ZM(\hat{\mathbf{F}})$, it is denoised through the Wiener filtering operation in the frequency domain as follows:

$$\begin{aligned}\tilde{\hat{\mathbf{F}}} &= WF\left(ZM\left(\hat{\mathbf{F}}\right)\right) \\ &= \mathscr{F}^{-1}\left\{\mathscr{F}\left(ZM\left(\hat{\mathbf{F}}\right)\right) - W\left(\mathscr{F}\left(ZM\left(\hat{\mathbf{F}}\right)\right)\right)\right\}\end{aligned} \tag{2.16}$$

where W denotes 3×3 Wiener filter whose variance is sample variance of the Fourier transform magnitude of $ZM(\hat{\mathbf{F}})$, i.e., $|\mathscr{F}(ZM(\hat{\mathbf{F}}))|$. It should be highlighted that all the matrix operations in this subsection are performed on element-wise basis.

2.2.3. PRNU estimation for video

Traditional frame based approach

PRNU noise estimate for a given video is performed very similarly to the reference PRNU estimation from still images as in Equation 2.16. The only difference is this time video frames are used instead of images. For this context, although all the video frames can be used, exploitation of a certain number, or a certain type of frames may also be sufficient for computation of a good quality PRNU noise from the video.

Video frames are encoded in 3 different kinds of frames, namely I, P, and B. P frames are encoded depending on the I or P frame that precedes it, whereas B frames are encoded

based on the P frame that precedes and that comes after it. I frames are encoded independently of P and B frames and the compression ratio is much lower than that in P and B frames. Thus, use of I frames in PRNU noise estimation provides more accurate pattern than that obtained from the same number of P and B frames. However, since I frames are usually encoded once or twice per second, using only I frames in short videos may be inadequate for estimating the PRNU with high accuracy (Chuang et al. 2011). The downside to using all frames is that it takes much longer time to process.

Block based approach

The video compression format H.264/AVC (Advanced Video Coding) is today's most commonly used video compression standard worldwide. It is based on block-oriented, i.e., each video frame is split into small blocks on each of which a set of operations are performed individually. In the prediction stage, a block's pixels are predicted from the encoded block(s) of the current frame, and/or of previous or future frames. The residue of the prediction (the difference between the actual block and the expected block) is then computed. In the transform stage, DCT (Discrete Cosine Transform) operation is performed on the prediction residue, followed by quantization and entropy coding.

If the DCT-AC coefficients in the prediction residue of an encoded block are totally zero, then in the decoding process, the high frequency content, and hence the PRNU noise in this block is displaced by that in its reference block(s) used for the prediction. Conversely, if the DCT-AC coefficients are not all zero, then the PRNU content in the block is preserved. Hence, the number of the non-zero DCT-AC coefficients in a block is a crucial factor affecting the strength of the remaining PRNU noise in the block.

The block-based PRNU estimation approach is based on detection of blocks whose prediction residue contains at least one non-zero DCT-AC coefficient, where PRNU noise is not entirely degraded, in each separate frame along the video. Only these blocks of the frames are used for computation of PRNU noise estimate. The others are discarded. For this purpose, a binary frame mask for each single frame is built. In a frame mask, the pixels locations of blocks whose PRNU content are totally degraded are assigned with zeros,

and the others are with ones. A pixel location (x,y) of the i th frame mask is obtained as follows:

$$M^{(i)}(x,y) = \begin{cases} 0, & \text{if all DCT-AC coefficients of the block that } (x,y) \text{ belongs to are zero.} \\ 1, & \text{otherwise} \end{cases} \tag{2.17}$$

Accordingly, if at least one non-zero DCT-AC coefficient exists in a certain block of the i th frame, all the corresponding pixels of this block in the frame mask $\mathbf{M}^{(i)}$ is set to one. Otherwise, they are set to zero. The PRNU noise in the block-based approach is computed using these frame masks in order to eliminate the non-valuable PRNU trace. Accordingly, the block-based video PRNU noise $\mathbf{K}$ can be obtained via a modified maximum likelihood estimator below:

$$\mathbf{K} = \frac{\sum_{i=1}^{L} \mathbf{N}^{(i)} \mathbf{I}^{(i)} \mathbf{M}^{(i)}}{\sum_{i=1}^{L} (\mathbf{I}^{(i)} \mathbf{M}^{(i)})^2 + J} \tag{2.18}$$

where, L denotes the number of the video frames, $\mathbf{I}^{(i)}$ is the decoded i th frame frame, $\mathbf{N}^{(i)}$ is the estimated PRNU noise of $\mathbf{I}^{(i)}$, and $\mathbf{M}^{(i)}$ is the corresponding frame mask. All the operations in Equation 2.18 are element-wise. J is a ones matrix and is added to avoid the potential case of $0/0$ when $M^{(i)}(x,y) = 0, \forall i$.

Zero meaning and wiener filtering operations are also performed in the block-based approach after the exploitation of the modified maximum likelihood estimator, as in Equation 2.16.

2.2.4. PRNU based source camera identification

In order to detect or verify if a given image is recorded by a suspect camera, a similarity test is performed between the estimated PRNU noise $\mathbf{N}_t$ and the reference PRNU $\tilde{\tilde{\mathbf{F}}}_r$ of the camera. For the similarity test, NCC operator (Normalized Cross Correlation) can be used as follows:

$$\rho(x,y) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(\mathbf{N}_t(i,j) - \overline{\mathbf{N}}_t)(\tilde{\tilde{\mathbf{F}}}_r(i+x,j+y) - \overline{\tilde{\tilde{\mathbf{F}}}}_r)}{||\mathbf{N}_t - \overline{\mathbf{N}}_t||||\tilde{\tilde{\mathbf{F}}}_r - \overline{\tilde{\tilde{\mathbf{F}}}}_r||} \tag{2.19}$$

where $||.||$ represents the Euclidean norm, and x and y denote the shift parameters. Nevertheless, use of peak correlation coefficient alone is not a reliable metric to set a common threshold value for different camera sensors, and different resolutions. In this context, Peak to Correlation Energy (PCE) is the most robust and reliable metric for determining the decision threshold value, which can be computed as follows (Fridrich 2009):

$$PCE\left(\mathbf{N}_t, \tilde{\tilde{\mathbf{F}}}_r\right) = \frac{\rho_{peak}^2}{\frac{1}{|s|-|\varepsilon|}\sum_{s\notin\varepsilon}\rho_s^2} \tag{2.20}$$

where ρ_{peak} denotes maximum cross-correlation coefficient between $\mathbf{N}_t$ and $\tilde{\tilde{\mathbf{F}}}_r$. ε refers to the correlation coefficients within a small range around ρ_{peak}, s denotes all indexes of correlation coefficients, and $|s| - \varepsilon|$ represents the number of the correlation coefficients other than ε.

Source camera identification for video is similar to that for image. The only difference is that $\mathbf{N}_t$ is obtained by using multiple video frames similarly to the case in Equation 2.15 except for that video frames are used instead of images. It is followed by zero meaning and wiener filtering operations respectively as in Equation 2.16.

3. MATERIALS AND METHODS

During the PhD period, we have developed a number of novel methods. In this section, these techniques are presented clearly.

3.1. Superpixel-based ENF Estimation for Video

In this subsection, a superpixel based ENF extraction technique for video is presented (Vatansever et al. 2017). The traditional method in the literature that is based on the phenomenon of global shutter sampling mechanism (Garg et al. 2011), (Garg et al. 2013) uses all steady pixels per frame to obtain one sample of illumination. This approach is unable to work in videos captured by CMOS cameras using sampling system of rolling shutter. This is because each row is captured in distinct time instances. In this thesis, unlike from the traditional techniques, we introduce an ENF estimation procedure exploiting stable superpixels, rather than using all stable pixels of a frame. A superpixel region consists of a set of pixels having similar pixel characteristics. ENF can be estimated through steady pixels belonging to a small superpixel region along successive frames of a given video. Estimation of ENF from a small superpixel region allows the proposed technique to work not only in videos by CCD cameras but also in those captured by CMOS camera. The scope of the proposed method is videos whose frame rate is not a divisor or multiplier of nominal ENF regardless of the sampling mechanism the camera uses. The segmentation algorithm of SLIC (Simple Linear Iterative Clustering) (Achanta et al. 2010) is employed to calculate superpixel areas. Figure 3.1 shows superpixels of a sample segmented image by the SLIC algorithm. The fundamental concept of the suggested technique for ENF extraction is all pixels in a superpixel region is presumed to show uniform reflectance features.

The quantity of illumination received by pixel coordinates x, y of an imaging sensor at moment n can be expressed as (Gonzalez and Woods 2011):

$$I(x,y,n) = i_s(x,y,n) \cdot r(x,y) \tag{3.1}$$

where $r(x,y)$ is the quantity of reflected illumination and $i_s(x,y,n)$ denotes instant illumi-

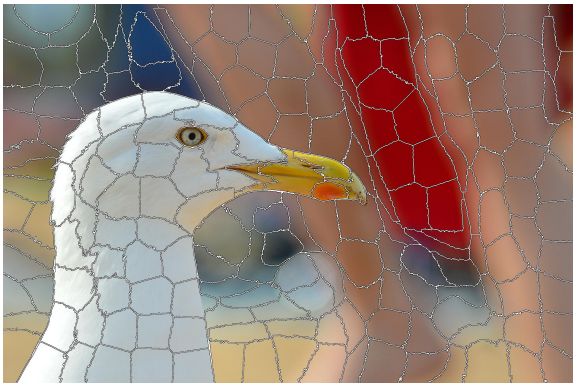


Figure 3.1. Superpixels of an exemplary segmented image

nation of a mains-powered source of light. A point light source, $i_s(x,y,n)$ can be roughly approximated in terms of electricity voltage as follows:

$$i_s(x,y,n) \approx \frac{\beta}{d_s(x,y)^2} \cdot \left|V(n)\right| \tag{3.2}$$

where $d_s(x,y)$ denotes the distance from the the light source to the spatial position (x,y), and β represents a transform factor to obtain luminance from voltage. The explanation of why $V(n)$ is in absolute form is a mains-powered bulb generates light in both negative and positive cycles of AC voltage. For any superpixel set, $r(x,y)$ can be considered as constant. Similarly, the distance from any pixel in superpixel set S to the light source $d_s(x,y)$ is also be assumed to be constant. Hence, for the k th steady superpixel set S_k, Equation 3.1 can be expressed as:

$$I_k(x,y,n) \approx \beta \cdot \frac{r_k}{d_k^2} \cdot \left|V(n)\right| \ , (x,y) \in S_k \tag{3.3}$$

where r_k represents the constant factor of reflectance for whole pixels in the k th steady superpixel region (superpixel set S_k), $k \in \{1,...,L\}$. L denotes the steady superpixels count, d_k is the distance from the light source to superpixel S_k. As can be noticed from Equation 3.3, $I_k(x,y,n)$ is directly proportional to $V(n)$. Accordingly, it indicates the frequency of mains electricity voltage (ENF) can be estimated from $I_k(x,y,n)$. For, any superpixel set S, illumination variations along video can be obtained by averaging the steady pixels in S for each frame individually, resulting in L illumination vectors in total. By utilization of any time or frequency domain techniques argued in (Garg et al. 2013), (Grigoras 2007),

ENF signal in video can be estimated through any of the illumination vectors. Here, we use STFT (Short Time Fourier Transform) based frequency estimation technique by using 20 seconds' data frames with one second shift, i.e., 19 seconds overlaps. Quadratic interpolation is then performed, consequently leading 1-second/sample ENF resolution. It is significant to note that the term "stable superpixel" or "steady superpixel" refers to the unchanged content of a superpixel set throughout the successive video frames. This will also be addressed in Section 3.2.

One weakness of this appproach is that it is unable to process videos whose frame rate are a nominal ENF divisor, such as 25 fps videos recorded in EU (50 Hz nominal ENF), and 30 fps videos captured in US (60 Hz nominal ENF). This is because, in this condition, the alias frequency is obtained at DC component.

3.2. Detecting the Presence of ENF in Video

In this subsection, an ENF presence detecting technique for digital video is introduced (Vatansever et al. 2017) based on superpixel based ENF estimation. The main operational steps of the proposed technique are illustrated in Algorithm 1. Accordingly, one frame, preferably the middle frame $\mathbf{F}_r$, in a selected shot C of a test video is segmented into superpixels. Superpixel segmentation is applied only to this representative frame. Then, in each superpixel region (superpixel set), the stable pixels, i.e. unchanged content, across all frames are located. Superpixels with a small amount of stable pixels ($m_l < \tau$) are disregarded. τ was empirically specified as 900 pixels. By using every consecutive frame $\mathbf{F}_n$ throughout the video shot, the average intensity $Y_k(n)$ is calculated individually from stable pixels of a superpixel set S_k. The same operation is repeated also for other superpixels. From each $\mathbf{Y}_k$ vector, a "so-called" ENF vector $\mathbf{E}_k$ is computed. The reason to exploit the "so-called" prefix is it is originally unknown whether or not it is actually ENF. Next, inter-similarity of the obtained ENF vectors is computed to determine if ENF exists in the given video. A representative vector of ENF $\mathbf{E}_r$ is calculated for this purpose first through element-wise median or mean operation performed on all computed vectors of ENF. Then, Pearson correlation coefficients between the representative vector $\mathbf{E}_r$ and

Algorithm 1. Proposed approach for detecting the presence of ENF

Step	Description
1	A video shot C is selected. Let $\mathbf{F}_n$ be the nth frame in C, where $n \in \{1,...,N\}$.
2	Middle frame in C is selected as the representative frame $\mathbf{F}_r$, where $r = \lfloor N/2 \rfloor$.
3	For the representative frame, superpixel sets are computed. Let Ω_l be the l th superpixel in $\mathbf{F}_r$, $l \in \{1,...,P\}$. P denotes the superpixels count.
4	Let Φ be the set of all stable pixels whose content are unchanged along C
5	Steady pixels count m_l in each Ω_l is computed by exploiting the stable pixel set Φ.
6	Stable superpixel set S is computed from $\{\Omega_l\}$: $S = \{\Omega_l \mid m_l > \tau\}$, where τ is a threshold value for m_l.
7	The average intensity $Y_k(n)$, $k \in \{1,...,L\}$, is computed for each stable superpixel S_k and each frame $\mathbf{F}_n$ from stable pixels only of region S_k. L denotes *stable* superpixels count.
8	Luminance variation signal $\mathbf{E}_k$ is computed from intensity variations $\mathbf{Y}_k$ for each superpixel S_k, individually.
9	Every $E_k(i)$ is placed into a matrix $\mathbf{M}$ such that $M(k,i) = E_k(i)$, where $i \in \{1,...,t\}$ and t is the length of the ENF vector. Let $\mathbf{E}_r$ be the representative ENF signal calculated via element-wise mean or median operation of all estimated $\mathbf{E}_k$ vectors, where nth sample of $\mathbf{E}_r$ is computed as: $E_r(i\|\text{mean}) = \underset{k}{\text{mean}}\{M(k,i)\}$ $E_r(i\|\text{median}) = \underset{k}{\text{median}}\{M(k,i)\}$
10	Similarity of each $\mathbf{E}_k$ to $\mathbf{E}_r$ is computed by Pearson correlation, $\rho(k)$.
11	Maximum, mean, median, and other similar statistics of ρ vector are computed as decision metrics.
12	The existence of ENF is confirmed if the resulted value of the exploited metric is higher than a decision threshold.

each estimated $\mathbf{E}_k$ are computed as follows:

$$\rho(k) = \text{corr}(\mathbf{E}_k, \mathbf{E}_r) = \frac{< \mathbf{E}_k - \bar{\mathbf{E}}_k, \mathbf{E}_r - \bar{\mathbf{E}}_r >}{\| \mathbf{E}_k - \bar{\mathbf{E}}_k \| \, \| \mathbf{E}_r - \bar{\mathbf{E}}_r \|} \tag{3.4}$$

where $< \cdot >$ denotes the dot product, and $\| \cdot \|$ is L_2 (Euclidean) norm. The sample mean is represented by overline. Afterwards, any of the following metrics is implemented for decision: $f_1 = \max(\rho)$, $f_2 = \text{mean}(\rho)$, $f_3 = \text{median}(\rho)$, $f_4 = \text{corr}(\mathbf{E}_i, \mathbf{E}_j)$, where $\mathbf{E}_i$ and $\mathbf{E}_j$ represent the vectors resulting in the greatest two values of ρ values, i.e. the top two vectors with the most similarity to $\mathbf{E_r}$. If the computed value of the decision metric is higher than a threshold value defined, the test video is decided to contain an ENF signal.

3.3. An Analytical Model for ENF Dependence on Idle Period

In this subsection the work of Su et al. (2014b) which is described in Section 2.1.7 is extended and an analytical model demonstrating how power of ENF, and how frequency of ENF change depending on idle period is presented (Vatansever et al. 2019a).

In Equation 2.10, F_m specifies the quantity of attenuation in the power of ENF signal depending on the ratio of L to M. Accordingly, by discounting F_m, the ω_y value making $\left|X\left(\frac{\omega_y L-2\pi m}{M}\right)\right|$ the same as $\left|X(\omega_0)\right|$ is the shifted angular frequency for a particular idle period, where ω_0 is nominal value of angular electric frequency. The following cases should be analyzed to find the value of ωy:

Case 1:

$$\left|X(\omega_0)\right| = \left|X\left(\frac{\omega_y L - 2\pi m}{M}\right)\right| \tag{3.5}$$

Accordingly;

$$\omega_0 = \frac{\omega_y L - 2\pi m}{M} \tag{3.6}$$

ω_y and ω_0 can be expressed as follows:

$$\omega_y = \frac{2\pi f_y}{F_r L} \quad \text{and} \quad \omega_0 = \frac{2\pi f_0}{F_r M} \tag{3.7}$$

where F_r is video frame rate, f_0 is nominal frequency of illumination (double the nominal ENF), and f_y is the emerging frequency of illumination. By substituting the equivalents in Equation 3.7, Equation 3.6 can be reformed as follows:

$$\frac{2\pi f_0}{F_r M} = \frac{2\pi f_y}{F_r L}\frac{L}{M} - \frac{2\pi m}{M}\frac{F_r}{F_r} \tag{3.8}$$

From Equation 3.8, for **Case 1**, f_y is yielded as:

$$f_y = f_0 + mF_r \tag{3.9}$$

where, $f_y < \frac{F_r L}{2}$ is obtained from the Nyquist theorem. $F_r L$ is the sampling rate. Accordingly, for **Case 1**, the range of m can be obtained from Equation 3.9 as:

$$m < \frac{L}{2} - \frac{f_0}{F_r}, \quad m \in \mathbb{W} \tag{3.10}$$

Case 2: By representing $X\left(\frac{\omega L - 2\pi m}{M}\right)$ as $\hat{X}(\omega)$, and considering the fact that Fourier transform is an even function, the following equivalents can be obtained:

$$\left|\hat{X}(\omega_y)\right| = \left|\hat{X}(-\omega_y)\right| = \left|X\left(\frac{-\omega_y L - 2\pi m}{M}\right)\right| \tag{3.11}$$

Similarly,

$$\left|X(\omega_0)\right| = \left|X(-\omega_0)\right| \tag{3.12}$$

Then, the following equivalent can be constructed as:

$$\left|X(-\omega_0)\right| = \left|X\left(\frac{-\omega_y L - 2\pi m}{M}\right)\right| \tag{3.13}$$

Accordingly,

$$\omega_0 = \frac{\omega_y L + 2\pi m}{M} \tag{3.14}$$

By substituting ω_0 and ω expressions in Equation 3.7, Equation 3.14 can be rewritten as follows:

$$\frac{2\pi f_0}{F_r M} = \frac{2\pi f_y}{F_r L}\frac{L}{M} + \frac{2\pi m}{M}\frac{F_r}{F_r} \tag{3.15}$$

From Equation 3.15, for **Case 2**, f_y results as:

$$f_y = f_0 - mF_r \tag{3.16}$$

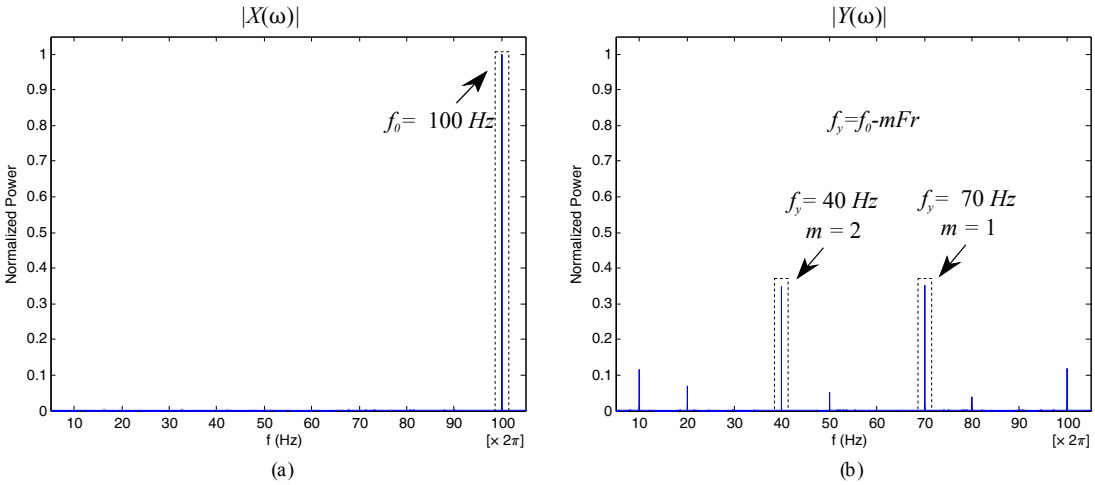


Figure 3.2. The frequency spectrum obtained for an 30 fps video recorded in EU, i.e, 50 Hz nominal ENF: (a) when there is no idle period, the ENF appears at nominal illumination frequency, i.e., 100 Hz; (b) when idle period of 45% is implemented, new ENF components are derived in certain frequencies - the estimated ENF power also decreases noticeably owing to the idle period

For **Case 2**, the set of m values for $f_y > 0$ can be obtained from Equation 3.16 as:

$$m < \frac{f_0}{F_r}, \quad m \in \mathbb{W} \tag{3.17}$$

It is important to highlight the range of m values provided in Equation 3.17 are for the f_y in Equation 3.16 only. Whereas, the range of m values in Equation 3.10 are for the f_y in Equation 3.9 only. Consequently, the following pair of equivalence can be formed:

$$f_y = \begin{cases} f_0 + mF_r, & \frac{f_0}{F_r} < m < \frac{L}{2} - \frac{f_0}{F_r}. \\ f_0 - mF_r, & m < \frac{f_0}{F_r}. \end{cases} \tag{3.18}$$

where $0 < f_y < \frac{LF_r}{2}$. The values of f_y leading to relatively high signal to noise ratio (SNR) can be utilized for a reliable ENF estimation. By exploiting the corresponding equivalences pair in Equation 3.18, the m value leading the greatest $\left|Y(\omega)\right|$ can be computed as:

$$m_p = \underset{m}{\operatorname{argmax}} \left| X_m \left(\frac{\omega L + 2\pi m}{M} \right) \times F_m(\omega) \right| \tag{3.19}$$

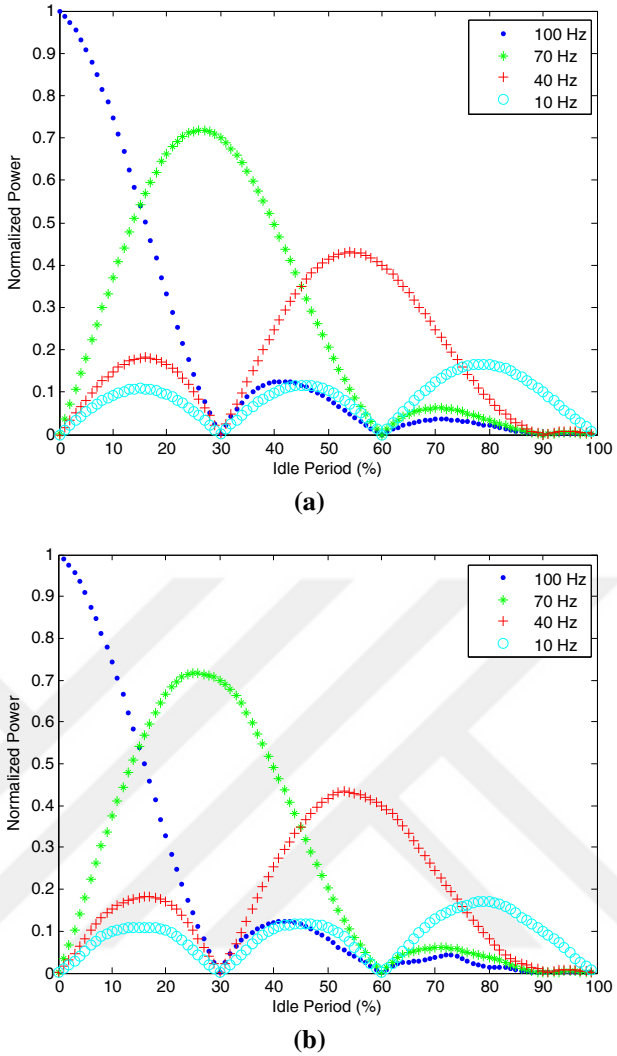


Figure 3.3. Frequency shift of primary ENF depending on the idle period for videos of 30 fps, and for a mains power network of 50 Hz: (a) the developed analytical model; (b) simulation

The corresponding f_y for the resulting m_p, which can be computed from Equation 3.18, is the greatest emerging component of ENF for a specific idle time, i.e., $\frac{M-L}{M} \times 100$. Figure 3.2 (a) depicts the Fourier spectrum for a 30 fps video recorded in the EU (50 Hz power network), given that the source camera does not apply any idle period for the specified settings. As the figure illustrates, the only component of ENF appears at 100 Hz, i.e., main illumination harmonic. Figure 3.2 (b) shows emerging components of ENF when an idle period of 45% is implemented by the camera. The greatest component of ENF appears here at 70 Hz, i.e., for $m = 1$, whereas the second greatest appears at 40 Hz, i.e., for $m = 2$. Interestingly, the power of both ENF components is very similar. This is because

the provided idle period is a point of transition for the strongest component. Figure 3.2 (b) also indicates that the captured ENF's power, i.e., normalized, significantly drops to 0.4 owing to the idle period. It is noteworthy that the other harmonics of ENF, in Figure 3.2 (a), and the other emerging components of ENF, in Figure 3.2(b), are discounted.

Figure 3.3 (a) shows the frequency variation of the strongest component of ENF depending on idle duration (in %) for a 30 fps video recorded in the EU. Accordingly, for 15%, 45%, 75% idle periods, the strongest component of ENF is replaced by respectively 70 Hz, 40 Hz and 10 Hz. It can also be seen in the figure that, as the idle duration increases, the captured ENF power decreases. These findings are also validated through simulation tests. As shown in Figure 3.3 (b), the outcomes of the simulation are almost identical to the proposed model findings. The minor variations are caused most likely by the additional components of ENF arising owing to the other illumination harmonics, 200 Hz, 300 Hz, etc. While there may be such additional components of ENF, both the proposed model and the simulation results validate that the greatest two components of ENF for any idle duration arises due to the 1^{st} illumination harmonic.

3.4. Proposed Idle Period Estimation Approach

In this subsection, a novel approach for idle period computation (Vatansever et al. 2019a) for ENF-containing videos exposed by rolling shutter is introduced. Although the methodology developed by Hajj-Ahmad et al. (2016), i.e. vertical phase method, may be effective for some cases, it has certain weaknesses. First, it relies on alias frequency, and is therefore not applicable to videos whose frame rate are a divisor of nominal ENF divisor, i.e., alias ENF arises at 0 Hz. Second, performing on noisy videos may be a challenge for their method owing to the difficulty in ENF phase estimation in such conditions. Our proposed approach can accommodate the limitations of their work.

The proposed approach is an extension of, and is based on the developed analytical model introduced in Section 3.3, and hence is targeted on videos captured by a rolling shutter system. According to the model, the 100 Hz harmonic is moved to a different frequency depending on the idle duration as depicted in Figure 3.3. As idle period duration increases,

Algorithm 2. Proposed idle period estimation method

Step	Description
1	Based on the nominal ENF of the connected mains power, and the frame rate of a given video, the analytical model for the variation of the ENF harmonic with the highest power, depending on idle period length, is computed by exploiting the technique introduced in Section 3.3.
2	Possible positions of ENF components in the test video is obtained from the derived model.
3	The time-series for luminance variation throughout the video is constructed. For this purpose, an illumination sample for each row is computed, followed by concatenation of the estimated luminance samples of all consecutive rows of all successive frames.
4	The strongest two components of ENF in the test video, H_1 and H_2, are located.
5	The power ratio of H_1 to H_2 is computed. Let the ratio be P_{H_1}/P_{H_2}.
6	By using H_1, H_2 and P_{H_1}/P_{H_2}, the corresponding range of the idle period is located on the analytical model illustration.
7	The middle of the located range is assigned as detected idle.

power of the second strongest component of ENF increases, whilst power of the strongest component decreases. The positions of these components, and the power ratio of them are the key features for the introduced technique.

The proposed approach consists a set of operational steps provided in Algorithm 2. Accordingly, reference analytical model for a given video is derived first based on on the video frame rate, and the nominal ENF as described in Section 3.3. Possible positions of ENF components in the test video is obtained from this model. Next, the time-series of luminance variation along the video is constructed by concatenating the row illumination samples of all successive frames. The reader is referred to (Su et al. 2014b) and (Wu et al. 2015) for the details on how to construct this time series. Discrete Time Fourier Transform is then computed for this series. The most powerful two components of ENF in the given video are computed. Then, the power ratio of the greatest component to the second greatest is calculated. Next, corresponding idle point in the analytical model is located based on the positions of these components, and the power ratio. Considering the noise factor, the candidate idle period range is allocated to a small band in the vicinity of the located point. This range's midpoint is accepted as estimated idle period.

As mentioned, the presented technique concatenates the luminance samples of consecu-

tive rows of every successive frame for construction of the luminance time series. Therefore, as discussed in Section 2.1.6, the sampling rate for the proposed technique is *number of rows* $\times$ *video frame rate* which is much greater than the Nyquist Criteria requires. Consequently, as alias ENF is not a case for the proposed approach, it can operate on videos at any frame rate, unlike the state of the art (Hajj-Ahmad et al. 2016) which relies on alias ENF.

The proposed algorithm's operational procedure is clarified in Section 4.1.4 with the experiments. Furthermore, it is illustrated with the experimental findings that how the proposed method can handle the state-of-the-art limitations.

3.5. Improved ENF-based Video Time-of-Recording Verification Method

An improved time-of-recording verification technique is proposed in this subsection for videos captured by the sampling system of rolling shutter (Vatansever et al. 2019a). It is mainly based on computation of potential ENF components arising due to the idle time, idle presumptions for each component, and the missing samples interpolation for each assumption.

In an ENF containing video, the strongest ENF components can be located based on the proposed analytical model introduced in Section 3.3. Nevertheless, in the frequency band of these components there may also be traces of non-ENF signals. This non-ENF signal could decrease the ENF signal quality to be extracted. On the other side, other components of ENF whose power are relatively lower may contain a good quality ENF. Thus, the proposed approach treats other components of ENF, particularly in the vicinity of the main illumination harmonic, as a potential source of a pure ENF signal carrier. It performs idle period assumptions, followed by missing samples interpolation for each presumption. These set of operations can result in higher quality ENF signal estimates, consequently. A better quality of ENF estimate generally results in a greater similarity, i.e., correlation coefficient, with the reference ENF, i.e., the ground-truth, and thus leads to better performance in time-of-recording verification. To our knowledge, no other technique in literature including (Garg et al. 2013), (Su et al. 2014b), (Wu et al. 2015) reap the benefit

Algorithm 3. Improved time-of-recording verification method

Step	Description
1	The time-series of variation of illumination throughout the test video period is computed by concatenating the row illumination samples of all successive frames as in Section 3.4
2	The potential components of ENF to be analyzed are identified based on introduced analytical model in Section 3.3.
3	The video idle period is assumed consecutively to be 5%, 10%, ... 95%.
4	For each assumption of the idle period, the missing luminance samples are interpolated for each frame by averaging the luminance samples of the next and the previous frame, consecutively leading the constructed time series to expand.
5	For the each reconstructed time-series, ENF vector is computed for each selected component of ENF.
6	By using NCC operator (normalized cross correlation), each computed ENF vector is compared with a reference ENF database, i.e., ground-truth. For each test of operation, the resulted peak correlation coefficient and and its lag point is recorded.
7	By using an appropriate metric, all computed peak correlation coefficients and their lag points are analyzed, and a final lag point is estimated depending on the exploited metric.
8	If the final lag point matches with the time period between the initialized time of the ground-truth and given time-of-recording of the video, the given time-of-recording is concluded to be correct.

of neither idle period assumption nor missing luminance samples interpolation for ENF extraction from videos exposed by the sampling mechanism of rolling shutter. Therefore, achieving ENF signal estimates of good quality may be a challenge for those techniques for some cases, such as videos with highly moving content, compressed videos, etc.

The proposed technique consists of a number of operational steps that are described in Algorithm 3. Accordingly, the time-series of luminance intensity variation along a given video period is constructed first through concatenation of the luminance samples of all successive rows of all consecutive frames as discussed in Section 3.4. Next, the potential ENF components that may arise due to the idle period are identified depending on the developed analytical model introduced in Section 3.3. The test video's idle period is assumed to be 5%, 10%, ... 95% consecutively. For each idle presumption, missing luminance samples in each frame are interpolated. In fact, estimating so many missing samples properly is a big challenge. Nevertheless, since the ENF signal does not indicate

sharp alterations, it is possible to make an approximation. Here, the exploited technique for interpolation is based simply on an average of luminance samples of the next and previous frame. ENF signal can be computed from each interpolated time-series with the use of any time or frequency domain methods provided in (Garg et al. 2013). Here, we prefer estimating ENF by means of STFT (Short Time Fourier Transform) based frequency estimation technique by using 20 seconds' data frames with one second shift, i.e., 19 seconds overlaps. Quadratic interpolation is then performed, consequently leading 1-second/sample ENF resolution. By using NCC operator (normalized cross correlation), each computed ENF vector is compared with a reference ENF database, i.e., ground-truth. For each test of operation, the resulted peak correlation coefficient and and its lag point is recorded. By using an appropriate metric, all computed peak correlation coefficients and their lag points are then analyzed, and a final lag point is estimated depending on the exploited metric. If the final lag point matches with the time period between the initialized time of the ground-truth and given time-of-recording of the video, the given time-of-recording, generally that provided in the meta data, is concluded to be correct.

4. RESULTS AND DISCUSSION

4.1. Experimental Work with ENF-based Media Forensics

In this subsection, experimental findings with the ENF based media forensics argued in the previous sections are provided.

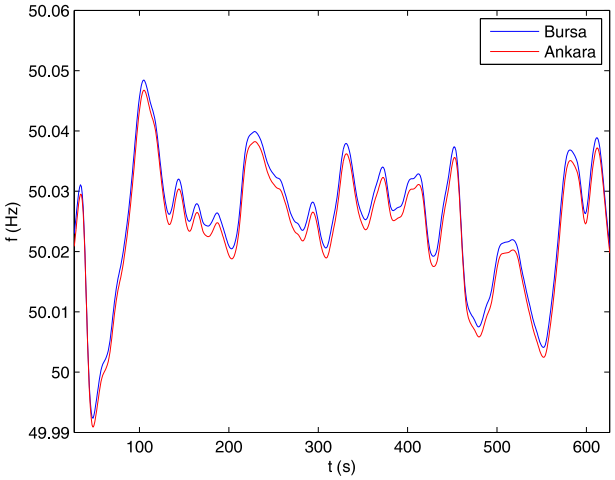
4.1.1. A study on Turkey's electricity system

As discussed in Section 2.1.1, the temporal variation of ENF in an interconnected power network is expected intrinsically to be identical at every point accross the entire network. Since the mains power grid in Turkey is interconnected, ENF at each point on the network should show the same oscillation (of the range 50 ± 0.1). In this context, the ground truth ENF data from the various cities of Turkey that are very far from each other were acquired as described in Section 2.1.4. It was observed that ENF variations obtained from these different cities are synchronous.

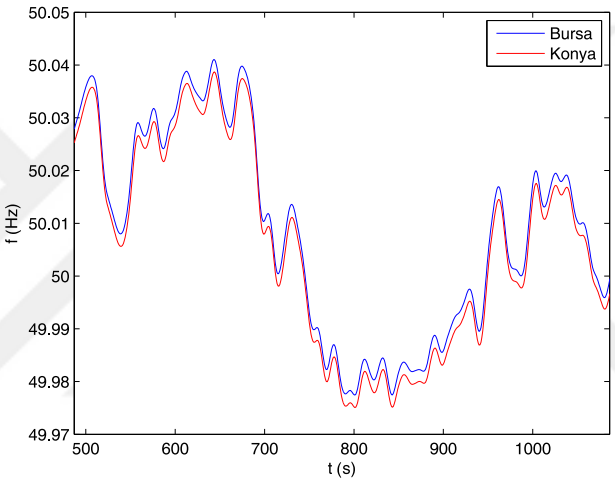
Figure 4.1 (a) shows that ENF data obtained in the same time of period from Bursa and Ankara are almost the same. Similarly, Figure 4.1 (b), and Figure 4.1 (c) show that ENF variations collected from Bursa and Konya; and those collected from Bursa and Kırıkkale respectively, each in the same time of period, are the same.

4.1.2. A comparative work of sources of ENF in audio and of microphone types

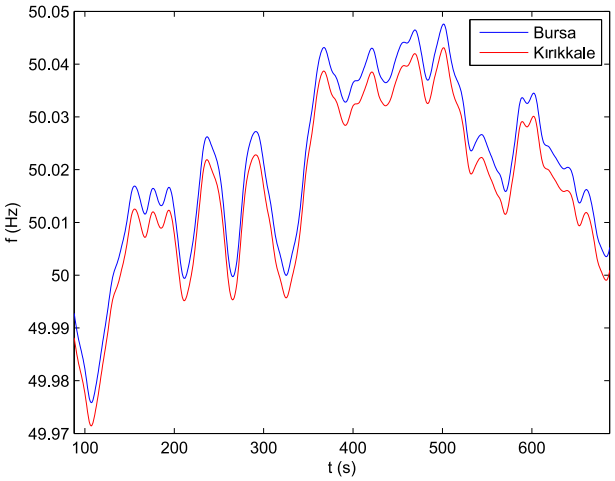
In this subsection, the effect of acoustic mains hum, and of electromagnetic propagation on ENF estimation from audio as well as the link between these sources and different microphone types on ENF are investigated. In this context, the first experiment was carried out on Viessman Vitopend 100 model boiler. In an indoor environment where all devices except for the boiler were switched off, audio recordings by a Samsung Galaxy S3 model mobile phone, equipped with electret microphone, and a dynamic microphone on an Asus K53S model laptop whose charger is plugged in were simultaneously initiated. The boiler was switched off 5 minutes after the audio recordings were started, but the recordings were continued for 5 minutes more, that is a total of 10-minute recording was



(a)

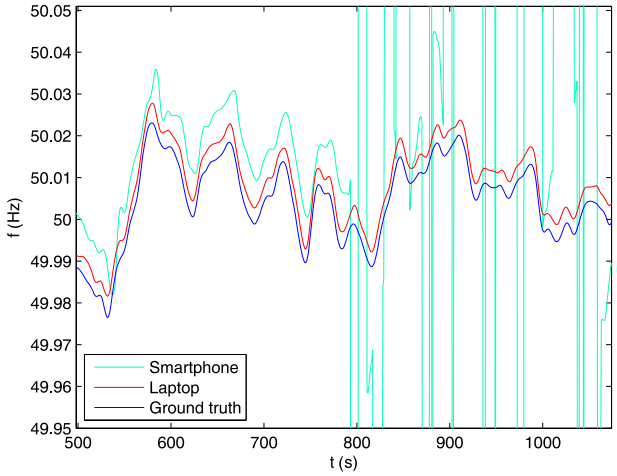


(b)

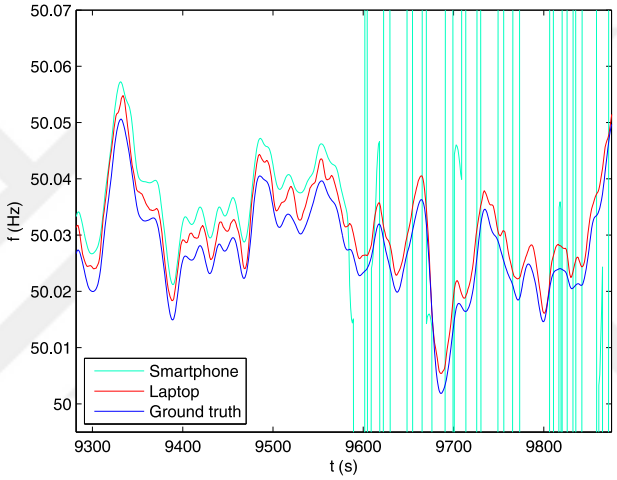


(c)

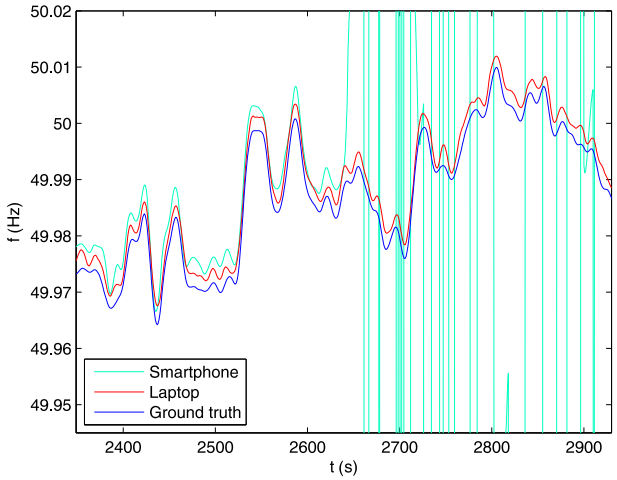
Figure 4.1. ENF signals acquired from different points in Turkey's interconnected power network: (a) Bursa vs. Ankara; (b) Bursa vs. Konya; (c) Bursa vs. Kırıkkale



(a)



(b)



(c)

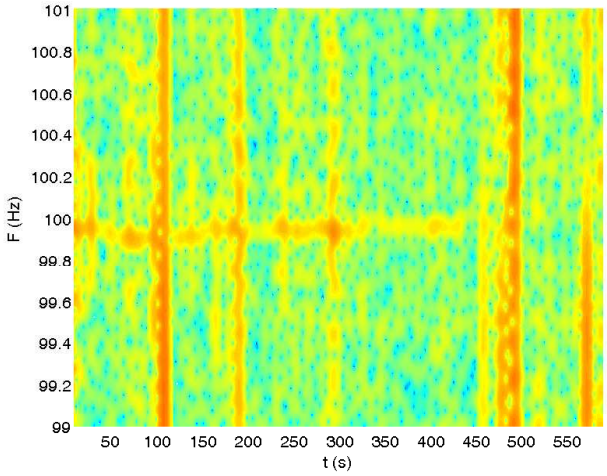
Figure 4.2. Household devices emitting acoustic mains hum: (a) boiler; (b) cooker hood; (c) vacuum cleaner

made by each device. The ENF signal in each recording was extracted by using the STFT based approach provided in Figure 2.1. Then, NCC operation was performed between the ground truth ENF database, i.e., reference ENF signal, and each estimated ENF signal. As a result of this operation, the value of peak correlation coefficient for each ENF signal was found in the 498^{th} sample of the reference ENF signal, which is depicted in Figure 4.2 (a). As ENF signals were estimated at 1 $sample/second$ resolution, it can be concluded that the audio recordings were started 498 seconds after the time the ground-truth ENF acquisition is initiated. It can also be yielded from Figure 4.2 that when the boiler is turned off, after the 783^{rd} seconds, the acoustic mains hum cannot be captured by the cell phone (turquoise line in the figure) as no acoustic mains hum is produced by any device from this point on. The presence of ENF in the dynamic microphone recording even after the acoustic mains hum generating device, i.e. boiler, is turned off is an indication that the source of ENF in the dynamic microphone recording is the electromagnetic field. As can also be seen from Figure 4.2 (a), the great similarity of the ENF signals obtained from each recording and the ground truth ENF are noteworthy.

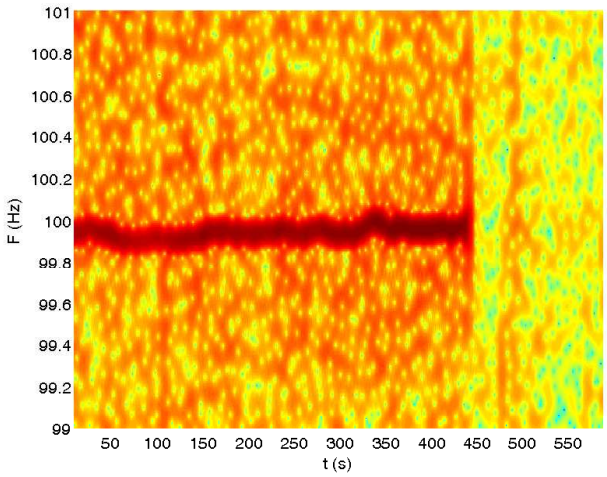
By using Arçelik P27i model cooker hood, and Arcelik S959 model vacuum cleaner as an acoustic mains hum emitting device, similar experiments were respectively performed. The yieldings obtained with these experiments are consistent with those obtained for the boiler as depicted respectively in Figure 4.2 (b) and Figure 4.2 (c).

As a result of the experiments conducted by using a number of different acoustic mains hum generating devices, including the aforementioned devices as well as air conditioners and hair dryers, the strongest, i.e. primary ENF harmonic sourced by acoustic mains hum is obtained around the 100 Hz frequency band. This is because the acoustic mains hum is generated in both negative and positive cycle of the electric voltage, hence it is double the frequency of mains voltage, i.e. double the ENF. Accordingly, each ENF signal estimated from each audio recording made by the mobile phone was divided by two so that visual comparisons with the ground truth ENF can be made around 50 Hz frequency band.

In this subsection, it is also investigated that if it may be possible to obtain a notable forensic information about the recording device in the sense whether it is equipped with



(a)



(b)

Figure 4.3. A comparison of acoustic mains hum interference into: (a) dynamic microphone; (b) electret microphone on spectrogram

electret or dynamic microphone, and about the recording scene in the sense whether there is an acoustic mains hum emitting device in the environment. It is mainly based on analysis of the primary frequency band ENF is detected, and the power of ENF at this band. The following experiment was conducted for this purpose.

Vieassman Vitopend 100 model boiler, which was previously verified to emit acoustic mains hum, was turned first on for 7 minutes, and then off for 3 minutes. In the meantime, in the same recording scene, a recording by the Samsung Galaxy S3 mobile phone (equipped with electret microphone), and another one by the Asus K53S laptop (with

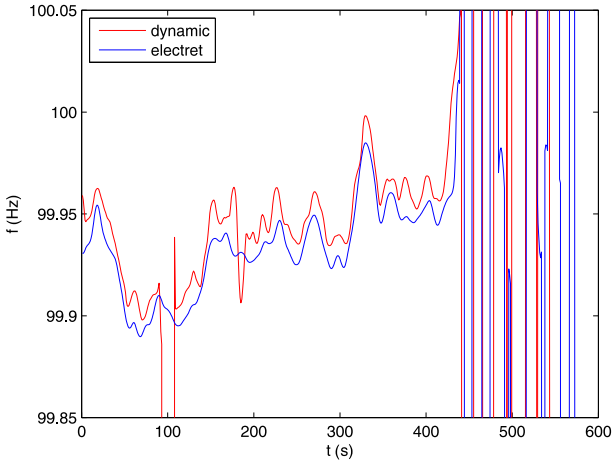


Figure 4.4. Estimated ENF signals from an audio recorded: (a) by dynamic microphone; (b) electret microphone - the only ENF source in the recordings is the mains hum

dynamic microphone) whose charge adapter was unplugged were made. The reason to the unplugged laptop charger is to reduce the electromagnetic field strength around the laptop, and thus around the dynamic microphone to a minimum. When the frequency spectra of the two audio were analyzed, it is seen that power of the ENF signals are very different from each other as demonstrated in Figure 4.3 (a) and (b). The frequency spectrum obtained from the dynamic microphone recording, and that from the electret microphone recording are respectively shown in the figure. Accordingly the 100 Hz band for the dynamic microphone, Figure 4.3 (a), is much weaker than that for the the electret microphone, Figure 4.3 (b). It is also obvious in Figure 4.3 (a) that the ambient noise, particularly around the 100^{th} second, is very high. This outcome is an indication that a dynamic microphones may not be applicable to analysis of and audio ENF whose source is acoustic mains hum. It should be highlighted that, the absence of the 100 Hz frequency band, i.e. ENF signal after the mains hum emitting device (the boiler) is switched off, in the last 3 minutes, illustrates that the boiler is the only source of the acoustic mains hum, and thus of the ENF signal. It may be considered that the source of ENF in the laptop recording may be electromagnetic field. However, no trace of ENF ,although not shown here, is detected at 50 Hz frequency band, which supports the outcome that the only source of ENF used for this experiment is the mains hum emitted by the boiler. Figure 4.4 depicts the estimated ENF signals from the dynamic microphone recording, and that from the electret microphone recording, which are obtained around the primary

frequency of the acoustic mains hum, 100 Hz. As can also be seen from the figure that there is no ENF trace for both dynamic microphone and electret microphone recordings after turning off the boiler. It is also noticeable how noisy the estimated ENF signal from the dynamic microphone recording is. It should also be noted that as both the ENF signal is obtained from the same frequency band, i.e., 100 Hz, the estimated signals were left as they are, i.e. they were not divided by two unlike to the previous experiments. Although not provided here, it may also be important to note that the power of ENF is obtained 1300 and 2.53, respectively for electret microphone recording and for dynamic microphone recording based on the square magnitude analysis in the frequency domain.

4.1.3. Evaluation of proposed superpixel based ENF presence detector

In this subsection, the efficiency of the proposed super-pixel based ENF presence detecting technique in Section 3.2 was assessed by conducting tests on videos with partially-moving content recorded in different outdoor and indoor environments in Turkey (50 Hz nominal ENF). A total of 160 videos were experimented for the presence of ENF signal, one half of which was captured with a Canon PowerShot SX210IS (CCD sensor) model camera, and other half captured with a Canon PowerShot SX230HS (CMOS sensor) model camera. The Canon SX230HS camera was deliberately selected as CMOS sensor since it is exploited in most video-ENF-associated studies (Garg et al. 2013), (Su et al. 2014a), (Su et al. 2014b), (Hajj-Ahmad et al. 2016). The reason to pick the Canon SX210IS is because it is the CCD equivalent of the SX230HS, and hence a reasonable comparison can be made for the sensor types of the same brand. One-quarter of the videos for each camera, were recorded at night. The exploited light sources for the illumination of recording scenes included tungsten, halogen, LED, fluorescent tube, CFL, and street light, i.e., all mains-powered, for this dataset. A second quarter were also captured under mains-powered sources of illumination, but this time in the daytime. Hence daylight is also included. The recording settings for this dataset included the environments such as a balcony, and a room with an open window. The third quarter were recorded in daylight only, i.e., non-main-powered source only. The last quarter were captured at night in the presence of non-mains-powered sources only, including vehicle headlight, moonlight, projection torch, candle, laptop screen, and smartphone torch. Accordingly, each

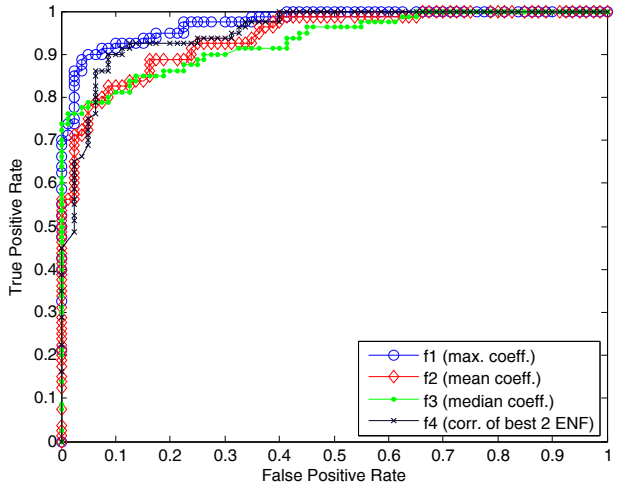


Figure 4.5. Assessment of the proposed superpixel based ENF presence detecting method when representative ENF vector is obtained via the *median* operation: ROC curves were obtained for videos half of which were captured by CCD sensor, and the others by CMOS sensors

Table 4.1. The performance of the proposed superpixel based ENF presence detecting method when representative ENF vector is obtained via the *median* operation: computed AUC values for the ROC curves

Sensor Type	# Videos	$f1$	$f2$	$f3$	$f4$
CCD	80	**0.985**	0.947	0.931	0.960
CMOS	80	**0.959**	0.942	0.941	0.944
Any (Mixed)	160	**0.973**	0.939	0.931	0.952

test video is known whether to include ENF. Each video was recorded at 29.97 fps in 640×480 resolution. During recording period of each video, the camera was fixed, and each video was ensured to have partially moving content. Peak alias frequency for 100-Hz signal is obtained at 10.09 Hz for 29.97 fps sampling frequency, i.e., frame rate, as mentioned in Section 2.1.6. Hence, the 10.09 Hz band of frequency was used to estimate the ENF signal from the test videos. The videos ranged from 2 to 15 minutes. However, for each video, a 2-minute-clip was exploited to search the presence of ENF. It is noteworthy that for each video the selected representative frame was segmented into superpixels in way that each superpixel region involves about 6400 pixels for a frame resolution of 640×480 pixels.

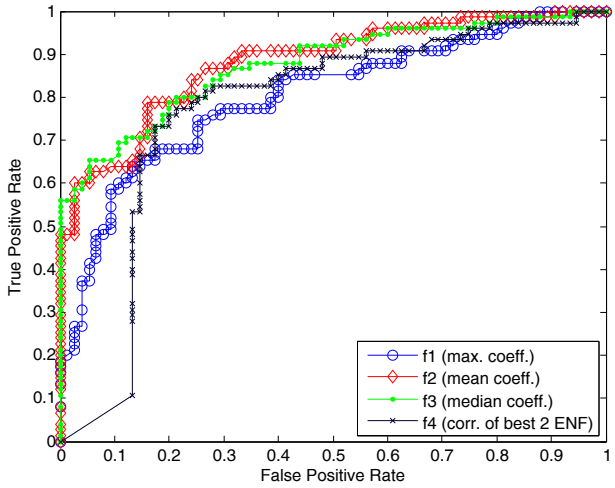


Figure 4.6. Assessment of the proposed superpixel based ENF presence detecting method when representative ENF vector is obtained via the *mean* operation: ROC curves were obtained for videos half of which were captured by CCD sensor, and the others by CMOS sensors

Table 4.2. The performance of the proposed superpixel based ENF presence detecting method when representative ENF vector is obtained via the *mean* operation: computed AUC values for the ROC curves

Sensor Type	# Videos	$f1$	$f2$	$f3$	$f4$
CCD	80	0.836	**0.896**	0.895	0.813
CMOS	80	0.760	**0.883**	0.866	0.700
Any (Mixed)	160	0.798	**0.886**	0.881	0.761

The performance of the proposed ENF detection technique was evaluated by computing an ROC (Receiver Operating Characteristics) curve and the area under this curve (AUC) for the introduced datasets. For this purpose, binary hypotheses below are introduced:

H_0 : ENF signal is absent in the given video.

H_1 : ENF signal is present in the given video.

Accordingly, the decision metrics f_1, f_2, f_3 and f_4, which are introduced in Section 3.2, were computed first for each video by exploiting both median-based representative ENF and mean-based representative ENF. Each computed value of each metric was assigned to one of the H_0 or H_1 cases.

Figure 4.5 provides the ROC curves constructed for each metric based on median based representative ENF estimation. Whereas Figure 4.6 illustrates ROC curves constructed based on mean based representative ENF. From the figures, considerable performance drop for all metrics is clearly seen when the decision metrics are computed using mean-based representative ENF. Table 4.1, and Table 4.2 show the AUC values (area under the curve) respectively for the ROC curves in Figure 4.5 and Figure 4.6. Resulted findings depending on the sensor type is also provided in these tables. The calculated values of AUC in Table 4.1 are significantly greater both for each sensor type and for the case of mixture. Here, f_1 metric surpasses the others.

4.1.4. Evaluation of proposed idle period estimation approach

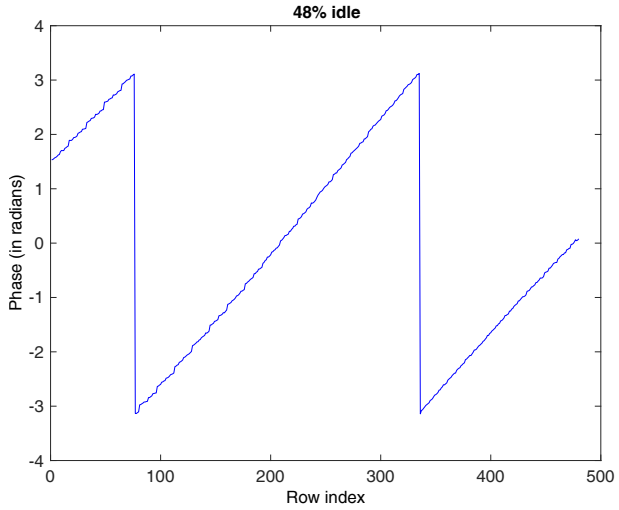
Videos with still-content

A comparative analysis is provided in this subsection on idle period estimation between an adaptation of the vertical phase analysis (Hajj-Ahmad et al. 2016), and our proposed method presented in Section 3.4. The results, along with some common idle-period properties, are demonstrated by experiments on still-conten (wall-scene) videos all of which contain ENF. Then, weaknesses of the vertical phase method, and how our proposed approach handles them are analyzed.

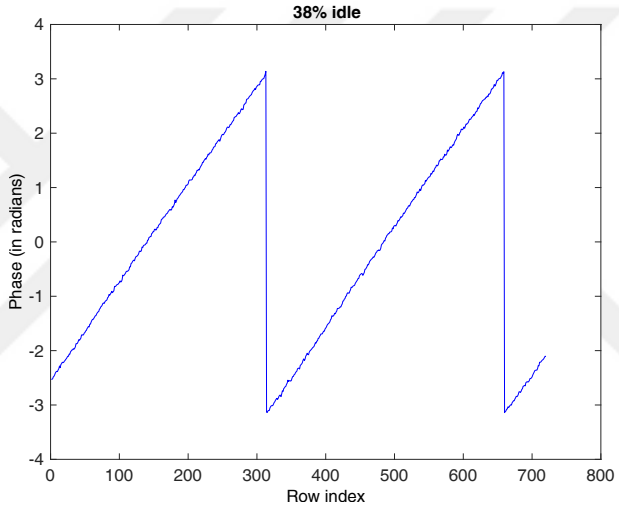
Vertical phase method can straightforwardly be adapted for idle estimation. The adapted method basically uses two key parameters: the sinusoidal cycles count obtained when idle period exists; and the cycles count that can be obtained in the absence of idle period. The proportion of these parameters constructs the key feature. Figure 4.7 (a) and Figure 4.7 (b) illustrates respectively estimated vertical phases for a 480p video (video-1), and a 720p video (video-2), both captured in Turkey at 30 fps by using a Canon SX230HS model camcorder. Idle period proportion R_{T_I} (in %) per frame can be obtained from the figures by using the following equation:

$$R_{T_I} = 100 - \frac{f_I}{F_r} \times \frac{100}{N_c} \tag{4.1}$$

where N_c is the sinusoidal cycles count obtained when idle time exists, which is equal



(a)



(b)

Figure 4.7. (a) Testing vertical phase method on a 480p video of wall-scene, video-1, resulting in an estimated idle period about 48%; (b) testing vertical phase method on a 720p video of wall-scene, video-2, resulting in an estimated idle period about 38% - both videos were captured in Turkey (50 Hz nominal ENF) at 30 fps by a Canon SX230HS model camcorder

the number of triangles (between $-\pi$ and $+\pi$) estimated via vertical phase analysis. f_I represents the nominal frequency of illumination, and F_r denotes video frame rate. $\frac{f_I}{F_r}$ provides the sine waves count that can possibly be obtained per frame in the absence of idle period. Accordingly the values of R_{T_I} for video-1 and video-2 were respectively computed as 48% and 38%. It can be deduced from these findings that idle period may change depending on video resolution at a fixed resolution.

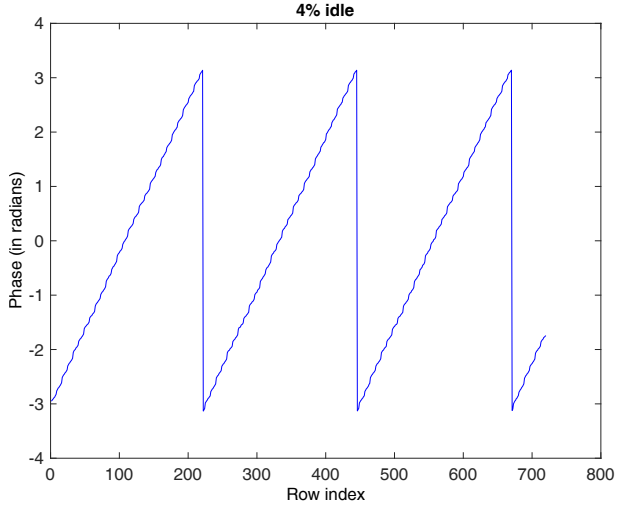


Figure 4.8. (a) Testing vertical phase method on a 30 fps video of wall-scene, video-3, resulting in an estimated idle period about 4%; (b) testing vertical phase method on a 23.976 fps video of wall-scene, video-4, resulting in an estimated idle period about 23% - both videos were captured by a Nikon D3100 model camcorder at 720p resolutions

Figure 4.8 (a) and Figure 4.8 (b) demonstrate estimated vertical phases for a 30 fps video (video-3), and a 23.976 fps video (video-4) respectively that are captured at 720p in Turkey by a Nikon D3100 model camcorder. The R_{T_I} values are estimated as 4% and 23% for video-3 and video-4 respectively. These findings indicate that another factor that affects idle duration is video frame rate.

Although vertical phase method may provide estimates of idle, it has certain weaknesses. First, estimation of vertical phase accurately may be challenging when ENF power in a

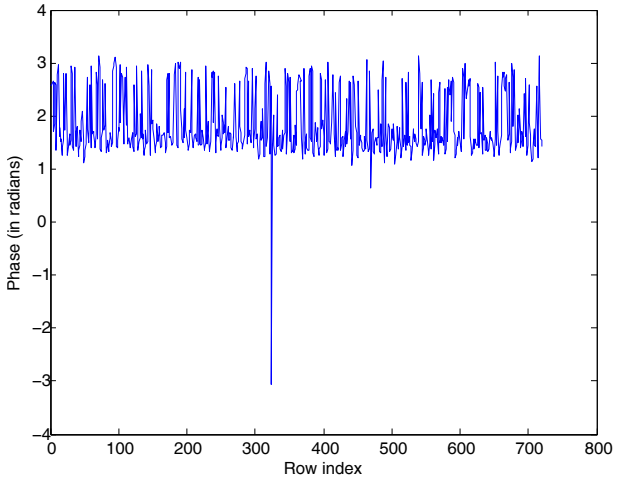
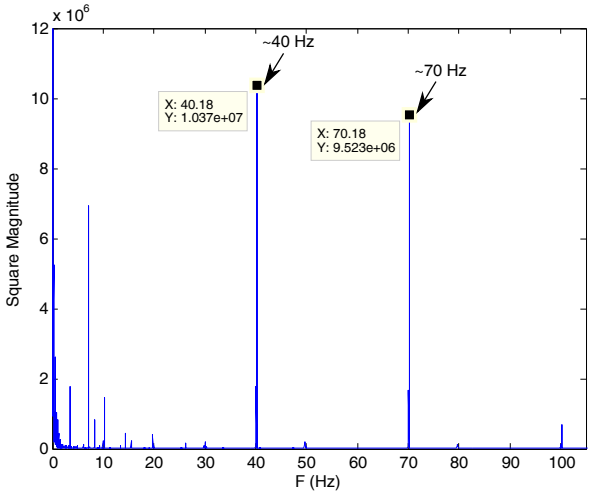


Figure 4.9. Vertical phase method performed for a 23.976 fps-720p video of wall-scene recorded in Turkey by the Nikon D3100 under CFL bulb illumination, video-5: computation of the phase is a great challenge for this video

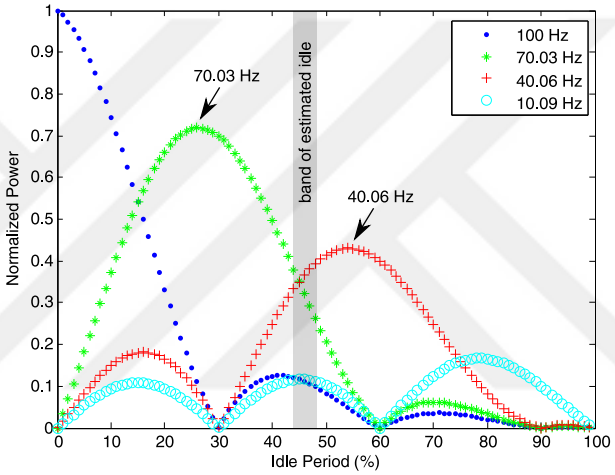
video is low owing to distinct reasons such as use of mains-powered light sources with unusual characteristics, long idle period, and high level of moving content. An illustration of such an example is provided in Figure 4.9 (a), which is obtained through analysis of a still content (wall-scene) video captured under CFL light. ENF phase estimation in such a condition is very challenging as can be deduced from the figure. It is notable that different kinds of mains-powered sources of light have different spectral responses, and CFL is one of those contributing the quality of ENF in video negatively (Vatansever and Dirik 2017), (Vatansever et al. 2019b). Another disadvantage of the vertical phase method is that it relies on alias frequency. Hence, processing videos captured by a frame rate value being a nominal ENF divisor is also a major challenge for this method. It should be recalled that in such a condition, alias ENF is computed at DC component, and making an approximation of the ENF from this DC element is extremely thorny.

The next experiments in this subsection present an assessment of our proposed approach on the same wall-scene videos addressed in this subsection. A comparison between the proposed approach and the vertical phase method is also made. Specifically, an exemplary video with 25 fps, i.e., a frame rate of nominal ENF divisor, is also tested. Vertical phase method cannot perform on such type of videos.

Fourier spectrum of the 480p video (video-1) captured by the Canon SX230HS at 30



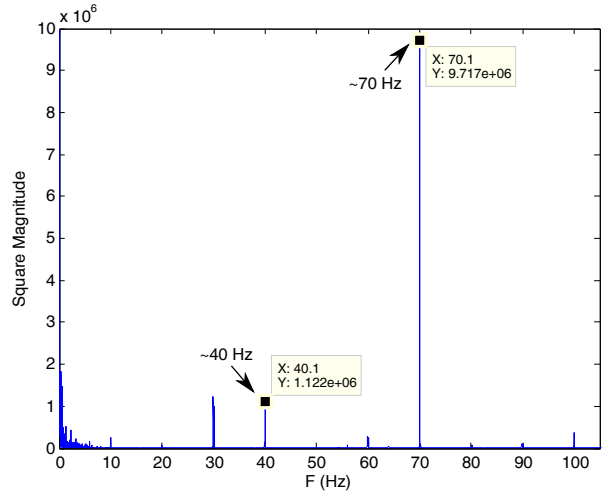
(a)



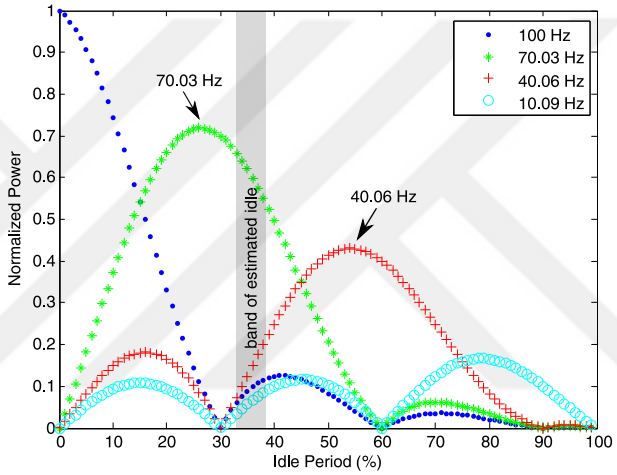
(b)

Figure 4.10. (a) Fourier Spectrum of the 30 fps-480p video captured by the Canon SX230HS, video-1; (b) constructed model of reference: variation in position of the primary ENF harmonic depending on idle duration in the same video - the range of idle duration was obtained between 45% and 50% (The measured idle by the vertical phase approach was 48%)

fps, and the reference model, constructed as introduced in Section 3.3, are respectively provided respectively in Figure 4.10 (a) and (b). It is clear in the Figure 4.10 (a) the ENF component with 40 Hz is the greatest, whilst that with 70 Hz is the second greatest. The ratio of the greatest component to the second greatest is about 1.1, which results in a possible idle period range between 45% and 50% as demonstrated in Figure 4.10 (b). Referring back to Figure 4.7 (a), idle duration was obtained 48% by vertical phase method for this video.



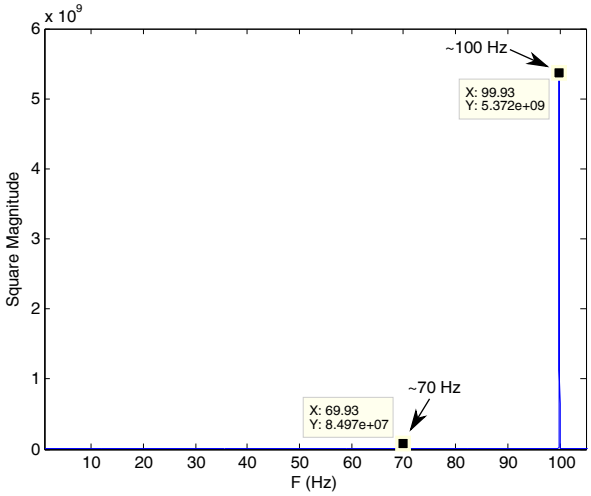
(a)



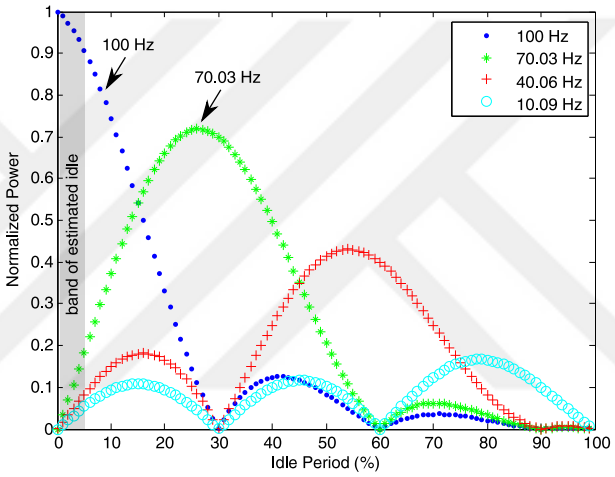
(b)

Figure 4.11. (a) Fourier Spectrum obtained for the 30 fps-720p video captured by the Canon SX230HS, video-2; (b) constructed model of reference: variation in position of the primary ENF harmonic depending on idle duration in the same video - the range of idle duration was obtained between 35% and 40% (The measured idle by the vertical phase approach was 38%)

Fourier spectrum of the 720p video (video-2) captured by the Canon SX230HS at 30 fps, and the reference model are respectively provided in Figure 4.11 (a) and (b). It can be seen from the Figure 4.11 (a) that the ENF component with 70 Hz is the greatest, whilst that with 40 Hz is the second greatest. The ratio of the greatest component to the second greatest is about 8.5, which results in a possible idle period range between 35% and 40% as shown in Figure 4.11 (b). It is noteworthy that constructed reference model for video-1, Figure 4.10 (b), and video-2, Figure 4.11 (b), are identical since the introduced technique is independent of video resolution. Referring back to Figure 4.7 (b) idle duration was



(a)



(b)

Figure 4.12. (a) Fourier Spectrum obtained for the 30 fps-720p video taken by the Nikon D3100, video-3; (b) constructed model of reference: variation in position of the primary ENF harmonic depending on idle duration in the same video - the range of idle duration was obtained between 0% and 5% (The measured idle by the vertical phase approach was 4%)

obtained 38% by vertical phase method for this video.

Similar experiments were repeated for the 30 fps-720p video (video-3), and for the 23.976 fps-720p video (video-4), taken by the Nikon D3100. While the range of idle for the video-3 was obtained between 0% and 5%, the range for the video-4 was obtained between 20% and 25% as illustrated in Figure 4.12 and Figure 4.13, respectively. It should be highlighted that constructed reference models for video-3 and video-4 are unlike as illustrated in Figure 4.12 (b) and Figure 4.13 (b). This is because the introduced method

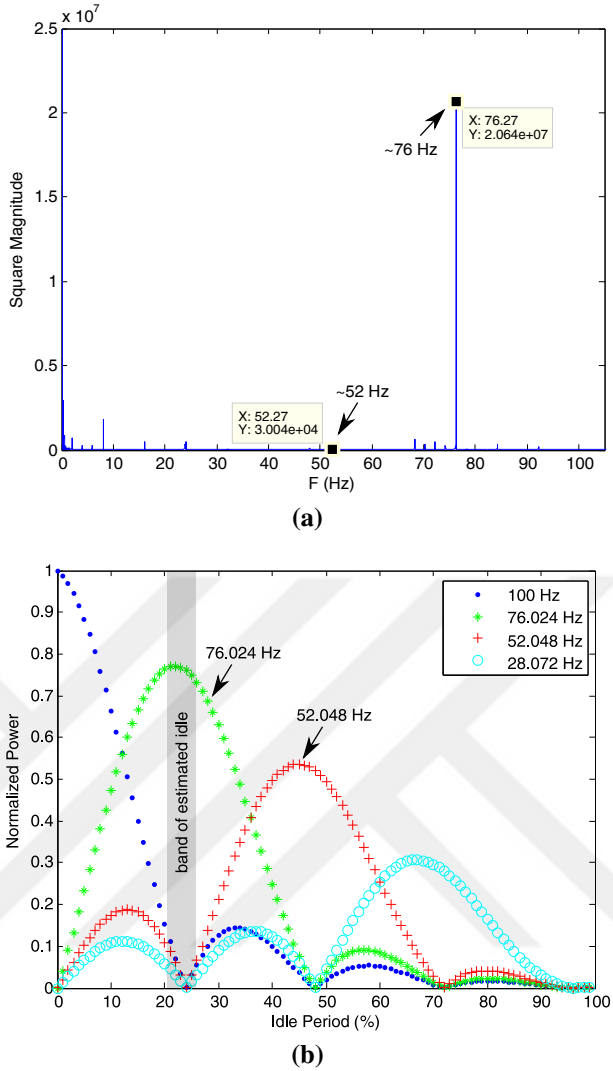
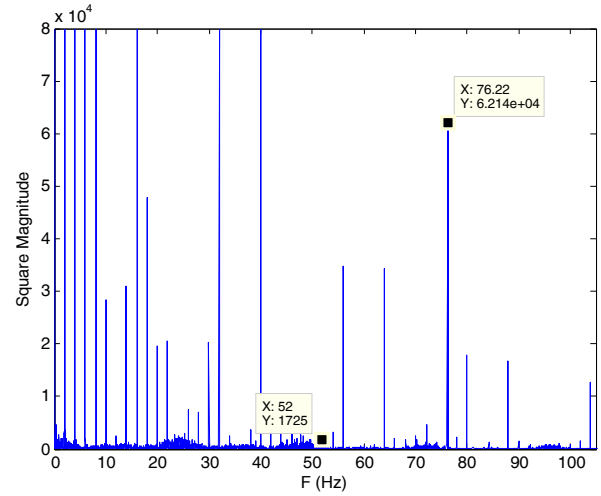


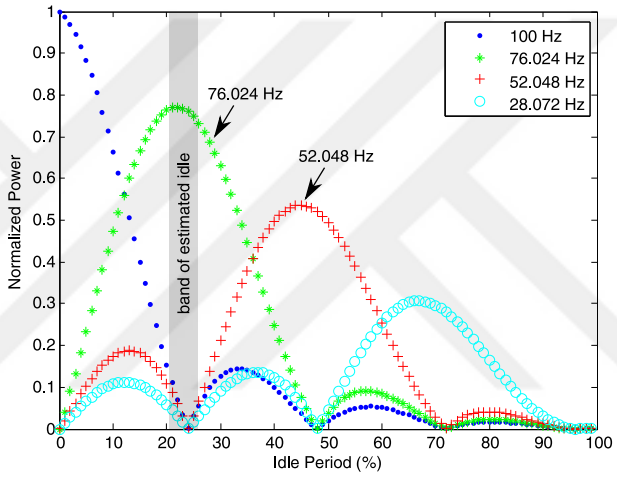
Figure 4.13. (a) Fourier Spectrum obtained for the 23.976 fps-720p video taken by the Nikon D3100, video-4; (b) constructed model of reference: variation in position of the primary ENF harmonic depending on idle duration in the same video - the range of idle duration was obtained between 20% and 25% (The measured idle by the vertical phase approach was 23%)

relies on the video frame rate. Returning to Figure 4.8(a) and Figure 4.8(b), idle duration for the video-3 and the video-4 were calculated 4% and 23% respectively via the vertical phase method.

Fourier spectrum of the other 23.976 fps-720p fps video taken under CFL light by the Nikon D3100 is provided in Figure 4.14 (a). For this video, the vertical phase method failed as discussed previously in this subsection, i.e., Figure 4.9. According to the reference model in Figure 4.14 (b), the idle duration range for this video was computed



(a)



(b)

Figure 4.14. (a) Fourier Spectrum obtained for the 23.976 fps-720p video taken under CFL light by the Nikon D3100, video-5; (b) constructed model of reference: variation in position of the primary ENF harmonic depending on idle duration in the same video the range of idle duration was obtained between 20% and 25% (The measured idle by the vertical phase approach was 23%)

between 20% and 25%, similar to that obtained for the video 4 as depicted in Figure 4.13. This finding is an indication that the suggested technique works more robust than the vertical phase in noisy videos in which the ENF power is relatively weak. This may be due to the fact that it is possible to maintain the power ratio between the greatest two components of ENF, even though the power of each is decreased.

In Figure 4.15 (a) and (b), frequency spectrum of a 25 fps -720p video (video-6), and the reference model is respectively shown. Based on the proposed approach, the idle

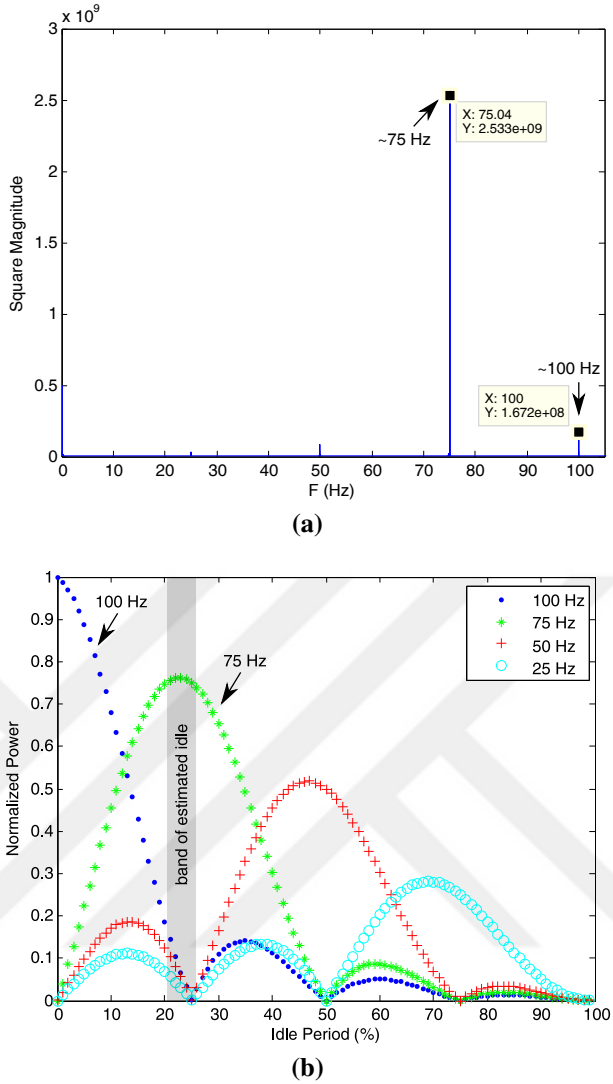


Figure 4.15. (a) Fourier Spectrum obtained for the 25 fps- 720p video taken by the Nikon D3100, video-6; (b) constructed model of reference: variation in position of the primary ENF harmonic depending on idle duration in the same video - the range of idle duration was obtained between 20% and 25% - alias ENF is detected at DC component for this frame rate, hence idle period for this case cannot be estimated via the vertical phase method

period range for this video was computed within the range 20% and 25%, similar to that obtained for 23.976 fps videos. However, in comparison to Figure 4.13 (a), the power ratio between the greatest two components of ENF is much smaller, which results in a lower idle duration. Accordingly, it may be deduced from this finding that idle duration is inversely proportional to video frame rate. This is because, a frame exposure time of videos at higher frame rate is expected to be smaller for a fixed resolution. Therefore, idle duration for video-6 is expected to be in closer proximity to 20% than the video-

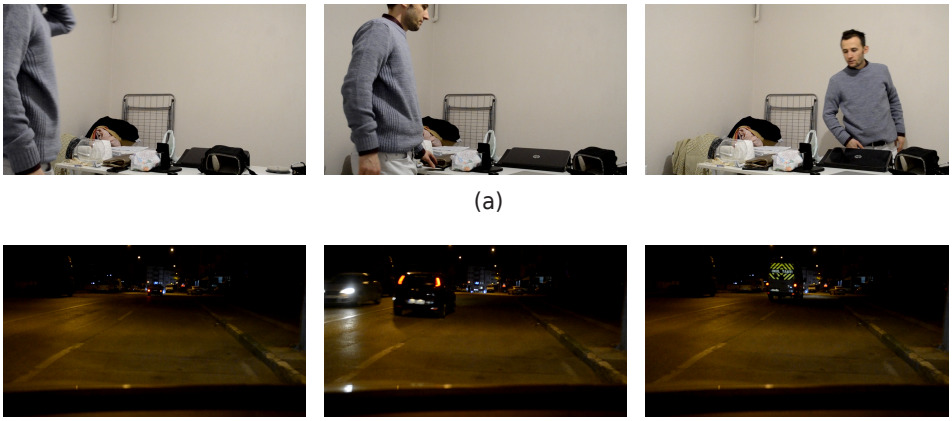


Figure 4.16. (a) Sample frames from an indoor video with moving content; (b) sample frames from an outdoor video with moving content

4. Accordingly, it can be concluded that the introduced approach works also on videos frame rate of which are a nominal ENF divisor. It may be considered although vertical phase analysis fails to estimate the idle period for videos at such a frame rate, processing a video at a frame rate of close proximity to the test video may lead an approximation to it. However, most consumer cameras do not offer sufficient number of options for frame rate settings. It should be highlighted that most camera manufacturers neither provide the idle period, nor frame read-out time. Hence, the performance of the introduced technique is assessed in both this subsection and next by benefiting from the vertical phase method.

Videos with moving-content

The introduced approach for idle period estimation is also assessed in this subsection by experimenting videos with moving content recorded by 5 different fixed cameras, i.e., videos of 5 different datasets. Each dataset contains six videos, one half are outdoor and the other are indoor. All of the videos were taken in Turkey (50 Hz nominal ENF). As power of ENF in a video, and hence the suggested technique can be influenced by the kind of light source, and the amount of luminous flux of the source, different settings were employed to create the datasets. Figure 4.16 (a) and (b) respectively provide sample frames from an indoor and outdoor videos.

Firstly, 2 video datasets by two different cameras were processed, and the estimated key

Table 4.3. Tabulated results for the computed reference model for the frequency shift of the main ENF harmonic depending on idle duration for a video captured at 29.97 in EU

Idle Period (%)	H_1 (Hz)	H_2 (Hz)	P_{H_1}/P_{H_2}
0	100	0	$\approx \infty$
5	100	70	5.0
10	100	70	2.0
15	100, 70	100, 70	1.0
20	70	100	2.0
25	70	100	5.0
30	70	100, 40	505.6
35	70	40	5.0
40	70	40	2.0
45	70, 40	70, 40	1.0
50	40	70	2.0
55	40	70	5.0
60	40	10, 70	498.5
65	40	10	4.9
70	40	10	2.0
75	40, 10	40, 10	1.0
80	10	40	2.0
85	10	40	5.1
90	10	40	334.0
95	10	40	6.9

parameters were provided for each video. The first dataset contain 480p videos recorded by the Canon SX230HS at 29.97 fps, and the other dataset have 720p videos taken by the Nikon D3100 at 25 fps.

Table 4.3 and Table 4.4 provide the greatest 2 ENF components and the ratio of power between them at idle periods of 5% increments for respectively 29.97 fps and 25 fps videos. Each entry in the tables was made by using the computed analytical model illustrations respectively in Figure 4.12 and in Figure 4.15. They form the reference parameters to compare with the estimated parameters obtained from the test videos. Table 4.5 provides the key information for each of test videos along with the estimated positions of the great-

Table 4.4. Tabulated results for the computed reference model for the frequency shift of the main ENF harmonic depending on idle duration for a 25 fps video recorded in EU

Idle Period (%)	H_1 (Hz)	H_2 (Hz)	P_{H_1}/P_{H_2}
0	100	0	$\approx \infty$
5	100	75	4.0
10	100	75	1.5
15	75	100	1.5
20	75	100	4.0
25	75	100	$\approx \infty$
30	75	50	4.0
35	75	50	1.5
40	50	75	1.5
45	50	75	4.0
50	50	100	$\approx \infty$
55	50	25	4.0
60	50	25	1.5
65	25	50	1.5
70	25	50	4.0
75	25	75	$\approx \infty$
80	25	50	6.0
85	25	50	3.5
90	25	50	2.7
95	25	50	2.2

est two components of ENF, and the computed power ratio between them. Also shown in Table 4.5 is the estimated idle period for each video, the expected idle duration, and the error. It should be re-emphasized that the camera manufacturers are unlikely to provide idle period information. Hence, the expected idle for 30 fps videos was obtained by exploitation of wall-scene videos analyzed in Section 4.1.4 via vertical phase analysis (Hajj-Ahmad et al. 2016). For the 25 fps dataset, the expected idle period was obtained in Section 4.1.4 based on the proposed method since vertical phase analysis method cannot operate in such a condition, i.e., a frame rate of nominal ENF divisor. The estimated idle period for each video was obtained based on detecting a match between computed parameters of reference model, and estimated parameters of a given video. Accordingly, for 10

videos, idle duration were computed in $\pm$ 5% vicinity of the expected idle as provided in Table 4.5. For 1 video, idle duration was computed in $\pm$ 10% vicinity of the expected idle, whilst for the other video the test failed.

Secondly, idle period estimation statistics for 5 different video datasets by 5 different cameras were analyzed. Each dataset consists of 6 videos at 30 fps and 720p, which were captured respectively by GoPro Hero 4, Nikon P100, Nikon D3100, Canon SX230HS and Canon SX220HS. Idle period for each video was estimated based on the proposed approach. The median value for all computed idle periods, and the average estimation error based on the computed median is provided in Table 4.6. Accordingly, the median value for each set of videos, i.e. for each camera, is obtained in very close vicinity of the expected idle. The average estimation error for GoPro Hero 4, for Nikon P100, and for Canon SX230HS is relatively smaller than that for Nikon D3100, and for Canon SX220HS. It can also be seen in the table that the idle duration for the videos by the Nikon P100, the Canon SX220HS, and the Canon SX230HS are very close to each other. Therefore, it can be concluded that similar or identical idle periods may be applied by different cameras.

In camera forensics, one potential task that idle period can be exploited is to verify whether or not 2 videos were created by different cameras. That is, as previously discussed, videos with similar idle periods may not suggest that they were taken by the same device. However, if the estimated idle periods of the query videos are quite different, it is most probable that the videos were originated from distinct cameras.

Table 4.5. The settings and computed findings for each processed video for idle period estimation based on the proposed technique

Camera Model	Vid. no	Res. (P)	Fr. rate (fps)	H_1 (Hz)	H_2 (Hz)	P_{H_1}/P_{H_2}	Estimated idle (%)	Expected idle (%)	Error (%)
Canon SX230HS	1	480	29.97	40	70	1.9	47.5	48.0	**0.5**
Canon SX230HS	2	480	29.97	70	40	1.0	45	48.0	**3.0**
Canon SX230HS	3	480	29.97	70	40	1.1	47.5	48.0	**0.5**
Canon SX230HS	4	480	29.97	40	70	1.27	47.5	48.0	**0.5**
Canon SX230HS	5	480	29.97	40	70	1.21	47.5	48.0	**0.5**
Canon SX230HS	6	480	29.97	40	70	1.4	47.5	48.0	**0.5**
Nikon D3100	7	720	25	75	100	3.5	17.5	21.5	**4.0**
Nikon D3100	8	720	25	75	100	3.6	17.5	21.5	**4.0**
Nikon D3100	9	720	25	75	100	2.1	17.5	21.5	**4.0**
Nikon D3100	10	720	25	50	100	1.6	No match	21.5	-
Nikon D3100	11	720	25	75	100	10.9	22.5	21.5	**1.0**
Nikon D3100	12	720	25	75	50	3.2	32.5	21.5	11.0

Table 4.6. The statistics for the idle period estimation for 29.97 fps-720p videos captured by different cameras based on the proposed method

Camera Model	Expected idle (%)	Estimated idle (%)	Avg. Estimation Error
GoPro Hero 4	72.69	72.50	1.66
Nikon D3100	4.30	5.00	4.16
Nikon P100	40.55	40.00	2.91
Canon SX230HS	38.00	37.50	2.50
Canon SX220HS	40.59	37.50	5.00

4.1.5. Experiments with improved time-of-recording verification

Videos with still-content

In this subsection, the introduced technique for time-of-recording verification, Section 3.5, was tested on a still-content, i.e., wall-scene, video dataset which was taken in Turkey by the Canon SX230HS and the Nikon D3100 camcorders. 3 different mains-powered light sources, namely CFL, LED, and Halogen bulbs were individually used for illumination of the scene. Under each of light sources, 2 videos with 29.97 fps in 720p were taken by each camera, leading to 12 native videos. Each of native videos was compressed in different bit rates (50k, 100k, 150k) via FFMPEG, and also YouTube. Accordingly, 48 compressed videos of wall-scene were obtained. The presented method was assessed by exploiting the below metrics:

Metric 1 (Garg et al. 2013): Peak correlation coefficient $(\rho_{H_i} > TH_c)$ at nominal frequency of illumination, $H = 100$ Hz. Idle period presumption is not performed, i.e., $i = \%0$.

Metric 2 (Su et al. 2014b): Supreme of peak correlation coefficients $(\rho_{H_i} > TH_c)$ computed for certain components of ENF, $H \in \{10, 40, 70, 100, 130, 160, 190, 200\}$ Hz. Idle period presumption is not performed, i.e., $i = \%0$.

Metric 3 (Proposed): Supreme of peak correlation coefficients $(\rho_{H_i} > TH_c)$ computed for certain components of ENF, $H \in \{10, 40, 70, 100, 130, 160, 190, 200\}$ Hz for 30 fps

Table 4.7. Assessment of the proposed method for time-of-recording verification on native wall-scene videos

Method	Metric 1	Metric 2	Metric 3	Metric 4
TD (%)	100	100	100	100
FD (%)	0	0	0	0
ND (%)	0	0	0	0

TD = True Decision, FD = False Decision, ND = No Decision

videos captured in EU. Includes idle period presumptions, $i \in \%\{0,5,10,\ldots,95\}$, and interpolation.

Metric 4 (Proposed): Supreme of normalized Euclidean distance (d_g) to $(n_{l_g}, \rho_{n_{l_g}})$.

$$d_g = \sqrt{{n_{l_g}}^2 + {\rho_{n_{l_g}}}^2} \tag{4.2}$$

n_{l_g} : lags count in a lag set. A lag set comprised of lags that are considered to be the same in a given tolerance. $\rho_{n_{l_g}}$: The supreme of peak correlation coefficients computed for each n_{l_g} $(\rho_{n_{l_g}} > TH_c)$. Metric 1 and Metric 2 are introduced on the basis of the work in (Garg et al. 2013) and in (Su et al. 2014b), respectively. Whereas Metric 3 and Metric 4 are suggested on the basis of the introduced technique in this thesis. It is notable that, based on empirical analysis, TH_c is selected as 0.94 .

The decision rates obtained for the native wall-scene videos are given in Table 4.7. In the table, ND refers to no decision, FD represents false decision rate, and TD denotes true decision rate. Accordingly, each metric performs 100% true decision rate. However, the performance considerably changes for the compressed videos as can be seen in Table 4.8. Proposed metrics, i.e., Metric 3 and Metric 4 show better performances than the others, and Metric 4 surpasses all. The true decision rate for Metric 4 is 70.83%, whereas it is 45.83%, 56.25% an 68.75%, respectively for Metric 1, Metric 2 , Metric 3.

Table 4.8. Assessment of the proposed technique for time-of-recording verification on compressed wall-scene videos

Method	Metric 1	Metric 2	Metric 3	Metric 4
TD (%)	45.83	56.25	68.75	70.83
FD (%)	0	4.17	4.17	2.08
ND (%)	54.17	39.58	27.08	27.08

TD = True Decision, FD = False Decision, ND = No Decision

Videos with moving-content

Time-of-recording verification approach was also evaluated in this subsection by using the same dataset as defined in Section 4.1.4, i.e., 24 videos with moving content. The metrics introduced in Section 4.1.5 were also used for these datasets to evaluate the effectiveness of the proposed technique. The only difference which is applied to the 25 fps videos only is that $\{25, 50, 75, 100, 125, 150, 175, 200\}$ Hz ENF components are used rather than $\{10, 40, 70,\ 100, 130, 160, 190, 200\}$ based on the analytical model introduced in Section 3.3. Except for that, all the steps are performed in the same way. Table 4.9 depicts the computed results for this dataset for each metric. Accordingly, Metric 3 and Metric 4 surpasses Metric 1 and Metric 2 for this dataset as well with 79.16% and 83.33% true decision rates, respectively. For Metric 1 and Metric 2, the true decision rates are 62.5% and 75%, respectively.

Experiments on both still-content videos and on videos with moving content illustrate that use of multiple ENF components, as well as idle period presumptions in each component and interpolation of missing luminance samples for each presumption, results better performance in time-of-recording verification than the state-of-the-art. Although Metric 3 and metric 4 results comparable outcomes, metric 4 outperforms.

4.1.6. Factors affecting ENF forensics in video

In this subsection, a number of factors that affect ENF based video forensics are investigated (Vatansever et al. 2019b). Firstly, the effect of type of mains-powered light source is explored. It is discovered that different kinds of illumination contribute to the quality

Table 4.9. Assessment of the introduced method for time-of-recording verification on videos with moving content

Method	Metric 1	Metric 2	Metric 3	Metric 4
TD (%)	62.50	75.00	79.16	83.33
FD (%)	8.33	8.33	12.50	8.33
ND (%)	29.16	16.66	8.33	8.33

TD = True Decision, FD = False Decision, ND = No Decision

of ENF signal to be estimated from video differently. Consequently, it affects the performance of ENF based video forensic tasks such as recording-time detection and verification. Secondly, it is studied how compression affects quality of ENF signal to be extracted from the video for different kinds of light sources. For this purpose, FFMPEG compression in different bit rates as well as a social media encoding, namely Facebook compression is investigated. Third, considering that ENF signal may show similar patterns in time, it is analyzed how ground-truth ENF database(reference ENF) of different lengths, with whom the similarity of the extracted video-ENF signal is compared, influences the recording-time detection and verification performance depending on video-ENF-signal length.

All experiments in this subsection were carried out using wall-scene videos recorded by a fixed camera. ENF signal estimations from these videos were made on the basis of averaging all the steady pixels in each frame throughout the video. That is, global shutter based approach discussed in Section 2.1.6 is used. The key metric used for detecting or verifying the recording-time of each video is the maximum correlation coefficient obtained as a result of normalized cross-correlation (NCC) operation between the estimated video ENF and the ground-truth. That is, the extracted video ENF signal is searched in a ground-truth ENF database, i.e., reference ENF signal, and the lag point at the peak coefficient computed by the NCC is recorded. If the time lag matches with the time difference between the actual recording-time and the initialized acquisition time of the ground-truth database, the video recording-time is verified. Otherwise, it is. These two cases correspond to the below binary hypotheses:

H0: The lag point at the peak NCC between the estimated ENF and the ground-truth ENF

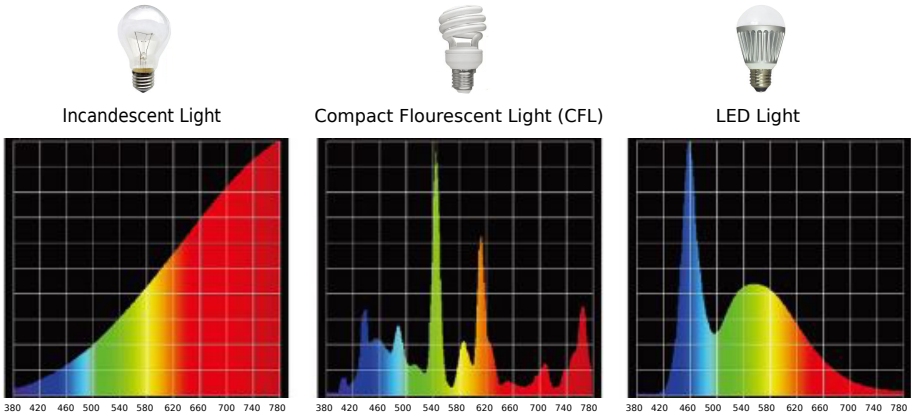


Figure 4.17. Spectra for different light sources

database, reference ENF, does not correspond to the time-of-recording.

H1: The lag point at the peak NCC between the estimated ENF and the ground-truth ENF database, reference ENF, corresponds to the video recording-time.

When an extracted video ENF signal is searched on a ground-truth database whose time span does not include the true recording-time of the video, i.e., presuming it is the false database, the time lag the resulted peak NCC coefficient inherently corresponds to H0 case. However, when searched on a ground-truth database whose time span does includes the true recording-time, i.e., assuming it is the true database, then the resulted time lag of the maximum NCC coefficient may correspond to either H1 case (if the time lag is a true positive, i.e., if matches with the true recording-time), or H0 case (if the time lag is a false positive, i.e., if does not matches with the true recording-time. To evaluate the detection and verification performance for each factor, ROC curves and the area under each curve, i.e., AUC are calculated. So as to constitute the ROC curve, each extracted vector of ENF is searched on both the correct database of the ground-truth, and a wrong database. Depending on the computed peak NCC coefficient and the time lag of this coefficient for each search, the corresponding case, i.e., H0 or H1 is determined and recorded.

The effect of type of illumination source

This subsection explores how the ENF quality in a given video is influenced by the form of mains-powered illumination source. Figure 4.17 depicts emission spectra of widely

Table 4.10. The light sources used in the experiments

Source No	Type	Color Temperature	Lum. Flux
S1	Halogen	2800 (K)	834 (lm)
S2	CFL	6500 (K)	870 (lm)
S3	CFL	2700 (K)	840 (lm)
S4	LED	6500 (K)	810 (lm)
S5	LED	2700 (K)	810 (lm)

used mains-powered sources, incandescent, white CFL, and white LED. It is clear in the figure that incandescent tungsten has a smaller level of blue light power, although it has the highest in red. Except for certain distinguishable peaks, like green and red, CFL has comparatively smaller level of power across the spectrum. LED emits the highest power in blue light, and comparatively smaller level of power in the red. Given these discrimination for different types of light sources across the visible spectrum, in the rest of this subsection, we investigate how the kind of light source can affect the quality of the estimated ENF signal, and consequently how it influences the application of ENF based time-of-recording detection and verification.

In this subsection, experiments were conducted by using five different kinds of light sources for illumination. Some specifications for these sources can be found in Table 4.10. According to the table, S2 and S3 represent white and yellow CFL respectively; S4 and S5 respectively denote white and yellow LED. S1 is Halogen, which emits yellow color. It should be higlighted that, although not included in Figure 4.17, there is almost no spectral difference between halogen land incandescent lamps.

For each light source in Table 4.10, 4 wall-scene (still-content) videos in about 10-minute length were captured at 30 fps at 480p by an SX210 IS model (CCD) Canon PowerShot camera. Each captured video was then split into 1, 2, 5, and 10-minute video clips resulting in 10, 5, 2 and 1 clips, respectively. Accordingly, for each kind of source of light, 40, 20, 8, and 4 clips in the 1, 2, 5, and 10 minutes duration were resulted. ENF signal was extracted from each clip. Then, NCC for each computed ENF vector was computed separately by exploiting both the true ground-truth ENF database of 24-hour length and a false

Table 4.11. Assessment of time-of-recording detection for light sources of different types: computed AUC values

Clip Length (min.)	S1	S2	S3	S4	S5
10	1.00	1.00	1.00	1.00	1.00
5	1.00	0.85	0.87	1.00	0.99
2	0.74	0.48	0.79	0.94	0.91
1	0.79	NA	0.32	0.45	0.65

Table 4.12. True recording-time estimations rate (%) for light sources of different types

Clip Length (min.)	S1	S2	S3	S4	S5
10	100	100	100	100	100
5	100	100	100	100	100
2	100	51.75	47.25	100	97.75
1	2.25	0.00	0.50	26.37	20.00

database of the same length, i.e., 24 hour. For each extracted ENF signal, the NCC procedure was repeated 20 times via shifting the initial time of the reference database 1 hour forward or back. That is, the initial time point and end point of the database were moved for each test. ROC curve was then computed on the basis of peak NCC distributions of H1 and H0 cases for each kind of source of light. Table 4.11 provides the corresponding AUC values for each ROC curve.

According to the table, yellow LED (S5) and white LED (S4) surpasses other sources of light for every video clip length. yellow CFL (S3) and white CFL (S2) performs comparatively worse. The reason to obtain small AUC values for small-length video clips, i.e., generally shorter than 2 minute, is incorrect matches caused by the similarity in variation of consecutive ENF samples in time. Longer ENF signals reduce the potential for the similarity of ENF pattern in time, and hence decreases false matches, resulting in higher AUC values. Table 4.12 shows the rate (in %) of true estimations of recording-time when extracted video-ENF vector is searched only in correct database of ground-truth. The findings in this table are similar to those reported in Table 4.11. As a result, different sources of light contribute differently to ENF in video. In videos captured under LED

Table 4.13. Assessment of recording-time detection for compression in different levels and type: computed AUC values for LED

Bit Rate	Average Size	10 min.	5 min.	2 min.
Original	938 MB	1.00	1.00	0.94
5000 Kbps	415 MB	1.00	1.00	0.94
1000 Kbps	94 MB	1.00	1.00	0.93
500 Kbps	53 MB	1.00	1.00	0.91
100 Kbps	20 MB	0.77	0.62	0.84
Facebook	13 MB	0.68	0.77	0.57

Table 4.14. True recording-time estimations rate (%) for compression in different levels and type for LED

Bit Rate	Average Size	10 min.	5 min.	2 min.
Original	938 MB	100	100	100
5000 Kbps	415 MB	100	100	100
1000 Kbps	94 MB	100	100	100
500 Kbps	53 MB	100	100	100
100 Kbps	20 MB	100	87.50	19.75
Facebook	13 MB	100	56.88	12.25

illumination, ENF signal is best estimated, whereas the estimated ENF signal quality for recordings made under CFL illumination drops significantly. All the experiments for the next two subsections were performed by exploiting only the sources of light which lead to the worst and to the best results, i.e. white CFL, and white LED.

The effect of compression

The compression effect on the ENF in video is explored in this subsection. Of the videos introduced in this subsection, those recorded under illumination of white CFL and white LED were compressed in H.264 compression standard with the use of FFMPEG at various bit rates; specifically 100 Kbps, 500 Kbps, 1000 Kbps and 5000 Kbps with the use of FFMPEG. Furthermore, each video was first uploaded, and then downloaded from Facebook. Accordingly Facebook compressed form of each video was built. For each form

Table 4.15. Assessment of recording-time detection for compression of different levels and type: computed AUC values for CFL

Bit Rate	Average Size	10 min.	5 min.	2 min.
Original	973 MB	1.00	0.85	0.48
5000 Kbps	430 MB	1.00	0.89	0.55
1000 Kbps	98 MB	1.00	0.89	0.66
500 Kbps	55 MB	1.00	0.66	0.11
100 Kbps	20 MB	0.98	failed	failed
Facebook	14 MB	failed	failed	failed

Table 4.16. True recording-time estimations rate (%) for compression in different levels and type for CFL

Bit Rate	Average Size	10 min.	5 min.	2 min.
Original	973 MB	100	100	51.75
5000 Kbps	430 MB	100	100	61.00
1000 Kbps	98 MB	100	100	42.50
500 Kbps	55 MB	100	100	6.00
100 Kbps	20 MB	51.25	0	0
Facebook	14 MB	0	0	0

of compression, the original video resolution was retained. Next, each compressed video was split into 2, 5 and 10-minute clips, and peak NCC coefficient distributions for the cases of H0 and H1 were computed for each clip, followed by construction of the ROC curves. AUC values obtained for LED case are provided in Table 4.13. Accordingly, the time-of-recording detection is achieved in considerably high performance up to a bit rate of 500 Kbps, almost for all size of video clips. The detection performance is significantly decreased by compression at 100 Kbps and Facebook. Table 4.14 demonstrates the true matches rate (in %), when the extracted video-ENF signals were searched only in the correct database of ground-truth ENF. According to the table, a similar performance is observed, except for 10-minute clips, where 100% achievement rate for compression of any kind is obtained.

Table 4.15 and Table 4.16 show AUC values and the correct matches rate (in %) respec-

Table 4.17. Assessment of recording-time detection for LED for ground-truth ENF data of different lengths: computed AUC values

Database Length	10 min.	5 min.	2 min.
One-day	1.00	1.00	0.94
One-week	1.00	1.00	0.91
One-month	1.00	1.00	0.82

Table 4.18. True recording-time estimations rate (%) for LED for ground-truth ENF data of different lengths

Database Length	10 min.	5 min.	2 min.
One-day	100	100	100
One-week	100	100	100
One-month	100	100	100

tively for CFL case. FFMPEG compression at 100 Kbps, and Facebook compression cause a complete failure in both detection (Table 4.15) and verification (Table 4.16) performances, for 5 and 2-minute clips, though, the performance is almost stable for compression with 1000 Kbps, and 5000 Kbps for each clip length. For 10-minute clips, Facebook compression again results in a failure, whereas 100 Kbps compression by FFMPEG leads an moderately acceptable performance with a 51.25% true detection rate (Table 4.16). The high value of AUC, 0.98, in the detection task for 100 Kbps compression in Table 4.15, is yielded since H0 and H1 cases are computed considerably distinct to each other, even though true matches count, i.e., H1, are only half of false matches, i.e., H0. Accordingly, the unexpected AUC values are clarified in the verification task by providing the true matches count.

The effect of length of ground-truth ENF

It is explored in this subsection that how the length of ground-truth ENF database, within which extracted video-ENF vector is searched, influences the recording-time estimation task depending on the extracted video ENF signal length. Of the videos defined in this subsection, those recorded under illumination of white CFL and white LED were split into

Table 4.19. Assessment of recording-time detection for CFL for ground-truth ENF data of different lengths: computed AUC values

Database Length	10 min.	5 min.	2 min.
One-day	1.00	0.85	0.48
One-week	1.00	0.80	0.45
One-month	1.00	0.75	0.38

Table 4.20. True recording-time estimations rate (%) for CFL for ground-truth ENF data of different lengths

Database Length	10 min.	5 min.	2 min.
One-day	100	100	51.75
One-week	100	100	27.50
One-month	100	98.13	11.00

2, 5 and 10-minute video clips. Then, the ENF vector for each clip was extracted, and each estimated signal was searched separately within true and false databases of ground-truth of one-month, one-week, and one-day length by using NCC operator. Each search was repeated 20 times by shifting the initial time of the database 24 hour, 8 hour, and 1 hours forward or back for one-month, one-week and one-day length reference ENF signal respectively. That is, for each test, the initial time point and the end point of the database were moved. Next, H0 and H1 cases were formed based on the peak NCC, followed by construction of the ROC curves. AUC values for LED cases are provided in Table 4.17. Accordingly, or 2-minute videos, the detection performance is in a trend of decrease as the reference ENF signal length increases. The detection performance for the 10-and 5-minute clips is completely stable for the ground-truth ENF database of each length, resulting AUC values of 1.00. True estimation rates (in %) of recording-time achieved when the extracted video-ENF vectors are searched only in the correct database is provided in Table 4.18. Surprisingly, 100% detection rate is obtained for every length of video clips and every length of reference ENF database.

AUC values and the rate of correct detection (in %) for CFL light are respectively given in Table 4.19 and Table 4.20. According to the tables, overall performance for 2-minute

videos reduces for both detection and verification as the reference ENF data length increases. The verification performance is considerably stable for every length of the ENF reference data, even though the detection performance drops slightly for 5-minute videos. The performance is stable for 10-minute videos, resulting in 100% estimation rate, for every clip length (10,5,2 minutes) and for every reference data length.

4.2. PRNU-based Source Camera Attribution for Social Media Video Pairs

In this subsection, a complementary analysis for block based source camera attribution proposed by Kouokam and Dirik (2019) is made by conducting experiments on social media videos, namely YouTube and WhatsApp. The main task that is targeted is whether two social media videos are originated from an identical source camera (no reference fingerprint is required for this task). For this purpose, PRNU noise estimate from both query video is computed by exploiting either I frames only, or all frames of the videos similarly to the case in Equation 2.16. Although I frames are more reliable for a good quality PRNU noise extraction from video, they are encoded in very small number compared to the sum of whole frames in a video, i.e., generally once or twice per second. Hence, using I frames only may be inadequate for some cases. But for some other cases, it provide a great advantage, particularly in computational time efficiency.

For the experimental work here, we use the non-stabilized social media videos in the "VISION" dataset introduced by Shullani et al. (2017). Table 4.21 provide the list of cameras and the set of non-stabilized YouTube videos exploited in the experiments, and Table 4.22 gives those of the WhatsApp. For each type of social media, the videos are divided into 2 distinct sets, namely flat videos, and natural videos. Flat videos set consists of wall-scene and sky-scene videos. Whereas, natural videos set is comprised of outdoor videos of natural scenes such as city, garden, etc., and of indoor videos of natural scenes such as offices, stores, classrooms etc. Each video has a length ranging from 65 to 75 seconds, and a sampling rate about 30 fps.

For the evaluation of the targeted task, the following binary hypothesis are exploited:
H_0: The tested two videos are captured by non-identical devices.

Table 4.21. The list of cameras and the set of non-stabilized YouTube videos exploited in the experiments

ID	Brand	Resolution	# flat	# natural	total
D01	Samsung Galaxy S3 Mini	1280×720	4	12	16
D03	Huawei P9	1920×1080	7	12	19
D07	Lenovo P70A	1280×720	7	12	19
D08	Samsung Galaxy Tab3	1280×720	13	21	34
D09	Apple iPhone4	1280×720	7	12	19
D11	Samsung GalaxyS3	1920×1080	7	12	19
D13	Apple iPad2	1280×720	4	12	16
D16	Huawei P9Lite	1920×1080	7	12	19
D17	Microsoft Lumia640LTE	1920×1080	4	6	10
D21	Wiko Ridge4G	1920×1080	4	7	11
D22	Samsung GalaxyTrendPlus	1280×720	4	12	16
D24	Xiaomi RedmiNote3	1920×1080	7	12	19
D26	Samsung GalaxyS3Mini	1280×720	4	12	16
D27	Samsung GalaxyS5	1920×1080	7	12	19
D28	Huawei P8	1920×1080	7	12	19
D30	Huawei Honor5c	1920×1080	7	12	19
D31	Samsung Galaxy S4Mini	1920×1080	7	12	19
D33	Huawei Ascend	1280×720	6	12	18
D35	Samsung Galaxy TabA	1280×720	4	12	16
Total			117	226	343

H_1: The tested two videos are captured by an identical device.

The computed PRNU noise for each test video is compared with the PRNU noise of other videos by using PCE operator as discussed in Section 2.2.4. The computed PCE value for each test is assigned either to the H0 case or to the H1 case. By using these H_0 and H_1 cases, ROC curve and are under the curve (AUC) are computed for each test for each method. H_1 cases are obtained by comparing a given video with other videos of the same device. Whereas, H_0 cases are obtained by comparing a given video with videos of all other devices separately. Hence the number of H_0 cases are always obtained much greater than that of H_1. The tests are performed mainly for 3 different different scenarios:

Scenario 1: The natural videos are compared with the natural videos, i.e.i natural vs. natural

Scenario 2: The flat videos are compared with the natural videos, i.e.i flat vs. natural.

Scenario 3: H_0 and H_1 cases of the *Scenario* 1 and *Scenario* 2 are combined.

Table 4.22. The list of cameras and the set of non-stabilized WhatsApp videos exploited in the experiments

ID	Brand	Resolution	# flat	# natural	total
D01	Samsung Galaxy S3 Mini	848 × 480	4	18	22
D03	Huawei P9	848 × 480	7	12	19
D07	Lenovo P70A	848 × 480	7	11	18
D08	Samsung Galaxy Tab3	848 × 480	12	21	33
D09	Apple iPhone4	848 × 480	7	12	19
D11	Samsung GalaxyS3	848 × 480	7	12	19
D13	Apple iPad2	848 × 480	4	12	16
D16	Huawei P9Lite	848 × 480	7	12	19
D17	Microsoft Lumia640LTE	848 × 480	4	6	10
D21	Wiko Ridge4G	848 × 480	3	7	10
D22	Samsung GalaxyTrendPlus	848 × 480	4	11	15
D24	Xiaomi RedmiNote3	848 × 480	7	12	19
D26	Samsung GalaxyS3Mini	848 × 480	4	12	16
D27	Samsung GalaxyS5	848 × 480	7	12	19
D28	Huawei P8	848 × 480	7	12	19
D30	Huawei Honor5c	848 × 480	7	12	19
D31	Samsung Galaxy S4Mini	848 × 480	7	12	19
D33	Huawei Ascend	848 × 480	7	10	17
D35	Samsung Galaxy TabA	848 × 480	4	12	16
Total			116	228	344

4.2.1. Experimental work on YouTube videos

In this subsection, a comparative analysis on PRNU based source camera attribution for YouTube video pairs was performed. The comparisons were basically made depending on the traditional video PRNU noise estimation, i.e., frame based approach, and block based approach, which are addressed in Section 2.2.3. For the block based approach, i.e., BB, all frames of the query videos were used to compute PRNU noise estimate. Whereas for the frame based method, two different PRNU noise estimates were obtained by exploiting I frames only, i.e., FB-I, and by using all frames, i.e., FB-all. As given in Table 4.21, YouTube videos dataset is comprised of 1080p and 720p videos. The comparisons were made for each set separately.

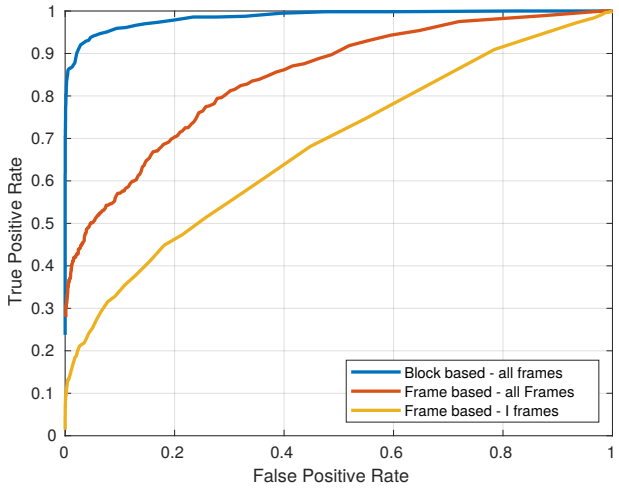


Figure 4.18. Computed ROC curves for all natural non-stabilized YouTube videos at 1080p in vision dataset - natural vs. natural

Table 4.23. AUC for All 1080p YouTube videos for different scenarios

Scenario	BB	FB-all	FB-I
Natural vs. Natural	0.99	0.84	0.68
Flat vs. Natural	0.98	0.91	0.86
Mixed	0.99	0.89	0.79

1080p videos

The performance analysis on the 1080p YouTube videos was first made by using all the videos of all cameras in this set. Figure 4.18 provides the computed ROC curves for the *Scenario* 1, i.e., natural videos vs. natural videos comparison. Accordingly, the block based method clearly outperforms the frame based approach for both the all frames case, and the I frame case. A similar test was also performed for the *Scenario* 2, i.e., flat videos vs. natural videos comparison, and also for the *Scenario* 3, i.e., a combination of *Scenario* 1 and *Scenario* 2. Table 4.23 shows the computed AUC values for each scenario. Accordingly, the block based method surpasses other approaches also for these scenarios, though the distinguishing performance was obtained for the *Scenario* 1. This is the expected result as the the block based approach is more effective in the changing content.

Table 4.24. AUC for each single 1080p YouTube video for 2 different scenarios, natural vs. natural, and flat vs. natural

	Natural vs. Natural			Flat vs. Natural		
Camera ID	BB	FB-all	FB-I	# BB-all	FB-all	FB-I
D03	0.99	0.85	0.62	0.94	0.92	0.87
D11	1.00	0.90	0.87	1.00	0.95	0.95
D16	0.98	0.81	0.68	0.96	0.95	0.85
D17	1.00	0.96	0.93	1.00	0.99	1.00
D21	1.00	0.89	0.79	1.00	0.88	0.96
D24	1.00	0.93	0.75	1.00	0.93	0.94
D27	1.00	0.82	0.65	0.98	0.92	0.91
D28	0.97	0.76	0.51	1.00	0.87	0.70
D30	0.95	0.74	0.57	0.96	0.73	0.66
D31	1.00	0.92	0.72	1.00	0.97	0.94

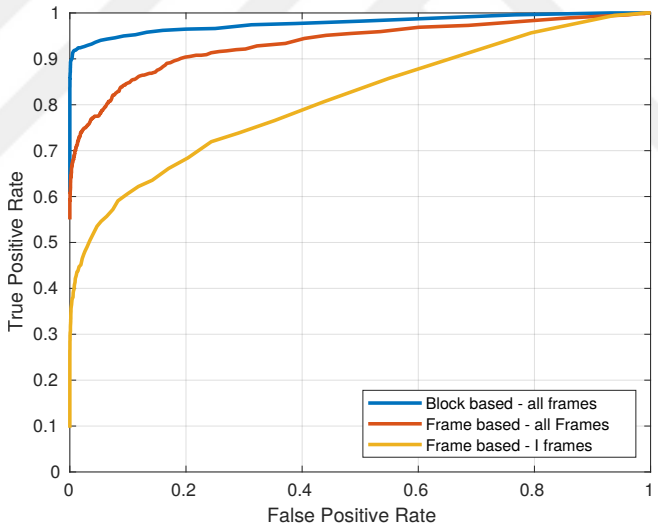


Figure 4.19. Computed ROC curves for all natural non-stabilized YouTube videos at 720p in vision dataset

As a second test, the performance analysis was made for videos of each camera individually. Table 4.24, provides the computed AUC values for each camera for each approach for *Scenario* 1 and *Scenario* 2. Accordingly, the block based method shows the greatest performance for each camera for each test.

Table 4.25. AUC for all 720p YouTube videos for different scenarios

Scenario	BB	FB-all	FB-I
Natural vs. Natural	0.98	0.93	0.81
Flat vs. Natural	0.97	0.97	0.93
Mixed	0.97	0.95	0.87

Table 4.26. AUC for each single 720p YouTube video for 2 different scenarios, natural vs. natural, and flat vs. natural

	Natural vs. Natural			Flat vs. Natural		
Camera ID	BB-all	FB-all	FB-I	# BB-all	FB-all	FB-I
D01	1.00	0.94	0.75	1.00	0.99	0.91
D07	0.82	0.73	0.61	0.74	0.83	0.70
D08	1.00	1.00	0.90	1.00	0.99	1.00
D09	0.98	0.95	0.95	1.00	0.99	1.00
D13	1.00	1.00	0.96	1.00	1.00	1.00
D22	1.00	0.98	0.79	1.00	0.99	0.93
D26	1.00	0.99	0.72	1.00	1.00	0.90
D33	1.00	0.96	0.69	1.00	0.99	0.86
D35	0.97	0.77	0.74	1.00	0.96	0.87

720p videos

Similar to the 1080p Youtube videos, the performance analysis on the 720p YouTube videos was made both for all videos of all cameras and videos of each camera individually. Figure 4.19 provides the computed ROC curves for all videos for the *Scenario* 1. Accordingly, the block based technique surpasses the frame based method for both the all-frames case, and the I-frames case. This test was also performed for the *Scenario* 2, and *Scenario* 3. Table 4.25 provides the AUC values calculated for each scenario. Consequently, the block based technique outperforms the others for these scenarios as well. The highest performance was achieved for also this video set in the *Scenario* 1.

Table 4.26, gives the obtained AUC values for each single device for each method for *Scenario* 1 and *Scenario* 2. Accordingly, the block based technique has the highest per-

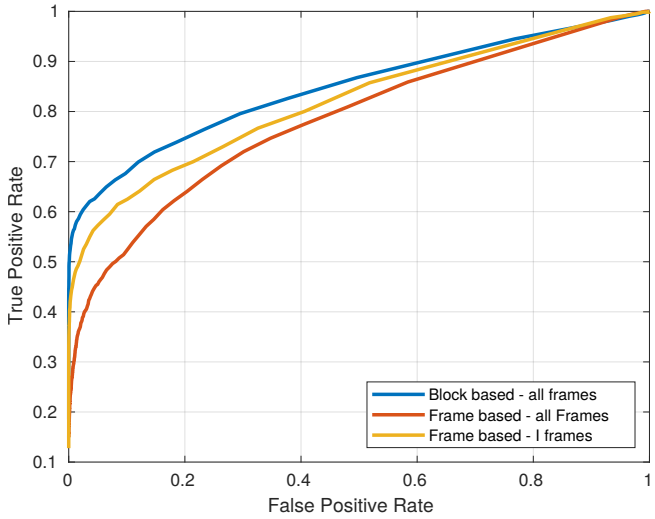


Figure 4.20. Computed ROC curves for all natural, i.e., non-stabilized WhatsApp videos in Vision dataset - natural vs. natural

Table 4.27. AUC for all WhatsApp videos

Scenario	BB	FB-all	FB-I
Natural vs. Natural	0.85	0.78	0.82
Flat vs. Natural	0.91	0.87	0.88
Mixed	0.88	0.83	0.85

formance for each test for almost each camera. The only outlier is obtained for the device D07, i.e., the Lenovo P70A in *Scenario* 2.

4.2.2. Experimental work on WhatsApp videos

In this subsection, a comparative analysis on PRNU based source camera attribution for WhatsApp video pairs was performed. The comparisons were performed in a similar way to that for Youtube videos as discussed in Section 4.2.1. As given in Table 4.22, WhatsApp videos dataset is composed of 480p videos.

For the performance assessment on WhatsApp videos, firstly ROC curves for the *Scenario* 1 was constructed for each method by using all the videos of all cameras. As illustrated in Figure 4.20, the block based method shows the greatest achievement. The test was repeated for the *Scenario* 2, and *Scenario* 3. Table 4.27 shows the computed AUC values

Table 4.28. AUC for each single WhatsApp videos for 2 different scenarios, natural vs. natural, and flat vs. natural

	Natural vs. Natural			Flat vs. Natural		
Camera ID	BB	FB-all	FB-I	# BB-all	FB-all	FB-I
D01	0.75	0.75	0.73	0.84	0.86	0.83
D03	0.67	0.69	0.61	0.72	0.79	0.68
D07	0.71	0.61	0.64	0.78	0.81	0.76
D08	0.92	0.89	0.90	0.98	0.96	0.98
D09	0.99	0.97	0.98	1.00	1.00	1.00
D11	0.94	0.79	0.88	0.96	0.86	0.89
D13	1.00	0.99	1.00	1.00	1.00	1.00
D16	0.66	0.67	0.66	0.85	0.86	0.77
D17	1.00	0.87	0.89	1.00	0.99	0.99
D21	0.88	0.73	0.75	0.95	0.83	0.91
D22	0.99	0.92	0.99	1.00	1.00	1.00
D24	0.88	0.79	0.84	0.98	0.80	0.94
D26	1.00	0.95	0.97	1.00	0.99	1.00
D27	0.78	0.70	0.77	0.91	0.88	0.89
D28	0.60	0.53	0.59	0.70	0.68	0.65
D30	0.65	0.61	0.62	0.74	0.65	0.66
D31	0.97	0.78	0.89	0.97	0.88	0.92
D33	0.97	0.87	0.93	0.92	0.89	0.93
D35	0.89	0.72	0.91	1.00	0.97	1.00

for each scenario. Accordingly, the block based method outperforms others for each scenario. Here, it is notable that for WhatsApp videos, frame based method performance for I-frames use only is greater than use of all frames.

Secondly, the performance analysis was performed for videos of each device separately. Table 4.28, present the calculated AUC values for each camera for each approach for *Scenario* 1 and *Scenario* 2. Although, the block based method shows the greatest performance for the majority of the tested cameras, there are a few outliers, i.e., D03 and D016 for the *Scenario* 1, and D01, D03, D07, D16, and D33 for the *Scenario* 2. These outliers may possibly be resulted from the low resolution of WhatsApp videos, i.e., 480p. That is,

the number of blocks used for PRNU noise extraction from WhatsApp videos of certain cameras may not be adequate for the block based method to work reliably. Either way, the block based approach for these devices performs worse only very slightly.

5. CONCLUSION

In this thesis, we have presented a comprehensive study on ENF (Electric Network Frequency) based multimedia forensics and introduced a number of novel approaches. We also provided a comparative analysis on PRNU (Photo Response Non Uniformity) based source camera attribution for social media video pairs.

Firstly, we provided a comparative research on the sources of ENF, i.e., electromagnetic field and mains hum, in audio. Accordingly, no electromagnetic field effect was observed in the audio recordings made by the electret microphone. On the other hand, the electromagnetic field interference was observed very high in audio recordings made by a dynamic microphone, yet the interference of the acoustic mains hum on this type of microphone recording was resulted to be very low. Accordingly, based on analysis of the primary frequency band of ENF, and of the power of ENF at this band, it may be possible to obtain a notable forensic information about the recording device in the sense whether it is equipped with electret or dynamic microphone, and about the recording scene in the sense whether there is an acoustic mains hum emitting device in the environment.

Secondly, we presented an ENF presence detecting technique for video. It is mainly based on the similarity of ENF vectors estimated from different superpixel regions. The motive to exploit superpixels for ENF estimation is that every pixel is almost uniform in brightness, texture and color in a superpixel region and therefore has uniform characteristic of reflectance. Utilization of such an area makes it possible to extract ENF signal from videos exposed by any sensor type, i.e., CCD or CMOS. The proposed method was tested on videos recorded by both CCD and CMOS sensors. Accordingly, for each sensor type, the proposed method performs considerably effective. It is notable that the proposed technique does not require any ground-truth ENF database. One weakness of the method is that it cannot process videos whose frame rate are a nominal ENF divisor, such as 25 fps videos recorded in EU (50 Hz nominal ENF), and 30 fps videos captured in US (60 Hz nominal ENF). This is because, in this condition, the alias frequency is obtained at DC component.

Thirdly, the phenomenon introduced by Su et al. (2014b) was researched further and an analytical model was developed to explore how mains-powered illumination frequency, hence the primary ENF, is attenuated and shifted depending upon idle period duration. This model led us a new method for idle period estimation to develop. The proposed method can handle the weaknesses of the traditional method such as its failure in videos at frame rate of a nominal ENF divisor and its inaccurate phase estimation in noisy videos whose ENF power is relatively weak. The proposed technique was tested on both videos with moving content and on wall-scene videos by different cameras. For each camera, the estimated idle was obtained in very close vicinity of the expected idle.

The aforementioned model also contributed to a new time-of-recording verification technique that is more effective than the traditional methods. A careful search for emerging components of ENF resulted from the idle period, followed by presumptions of idle period for each emerged component and interpolation for each presumption lead to more accurate ENF signal estimates. A more accurate ENF signal consequently results in higher efficiency in time-of-recording verification. The proposed technique was tested on both wall-scene compressed videos and on videos with moving content. In each case, the proposed method outperformed the traditional methods.

Another research provided in this thesis was the study on factors affecting ENF forensics for video. Accordingly, distinct sources of illumination, distinct compression ratios, and reference ENF signal (ground-truth) of different lengths affect the efficiency of the video time-of-recording estimation task noticeably. Particularly short videos are highly affected negatively from these affects. It can also be concluded that the primary factor affecting video ENF forensics is the type of mains-powered source of light. Best results for video-recording-time verification task was obtained for the videos taken under LED illumination. Whereas, the videos taken under CFL illumination yielded the worst performance based on the experimental analysis.

In this thesis, we also provided a comparative study on PRNU (Photo Response Non Uniformity) based source camera attribution for the social media videos in the "VISION" dataset introduced by Shullani et al. (2017). In this context, the block-based PRNU noise

estimation method proposed by Kouokam and Dirik (2019) was further studied, and a comparative and complementary analysis was provided via the experimental work on the non-stabilized YouTube videos and on the non-stabilized WhatsApp videos. Accordingly, block-based approach outperformed the traditional technique for the tested social media videos.

REFERENCES

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S. 2010. SLIC superpixels. EPFL Technical Report 149300.

Al-Athamneh, M., Kurugollu, F., Crookes, D., and Farid, M. 2016. Digital Video Source Identification Based on Green-Channel Photo Response Non- Uniformity (G-PRNU). Proceedings of the Fifth International Conference on Signal, Image Processing and Pattern Recognition (SPPR 2016).

Bayram, S., Sencar, H.T., Memon, N. 2009. An efficient and robust method for detecting copy-move forgery. IEEE International Conference on Acoustics, Speech and Signal Processing. 19-24 April, 2009, Taipei, Taiwan.

Bollen, M.H., Gu, I.Y. 2006. Signal processing of power quality disturbances. Wiley-Interscience.

Brixen, E.B. 2007. Techniques for the Authentication of Digital Audio Recordings. In Audio Engineering Society Convention 122.

Bykhovsky, D., Cohen, A. 2013. Electrical network frequency (ENF) maximum-likelihood estimation via a mul- titone harmonic model. *IEEE Transactions on Information Forensics and Security*, 8(5):744–753.

Chai, J., Liu, F., Yuan, Z., Conners, R.W., Liu, Y. 2013. Source of ENF in battery-powered digital recordings. Audio Engineering Society Convention 135.

Chen, M., Fridrich, J., Goljan, J. Lukas, M., 2007. Source digital camcorder identification using sensor photo response non-uniformity. Proc. SPIE 6505, Security, Steganography, and Watermarking of Multimedia Contents IX, 1 March 2007, San Jose, CA, USA.

Chen, M., Fridrich, J., Goljan, M., Lukas, J. 2008. Determining image origin and integrity using se*nsor noise. IEEE Transactions on Information Forensics and Security*, 3(1):74–90.

Chuang, W.H., Garg, R., Wu, M. 2013. Anti-forensics and countermeasures of electrical network frequency analysis. *IEEE Transactions on Information Forensics and Security*, 8(12): 2073–2088.

Chuang, W.H., Su, H., Wu, M. 2011. Exploring compression effects for improved source camera identification using strongly compressed video. Proceedings - International Conference on Image Processing, 11-14 September, 2011, Brussels, Belgium.

Cooper, A. 2008. The electric network frequency (ENF) as an aid to authenticating forensic digital audio recordings - an automated approach. 33rd International Conference: Audio Forensics-Theory and Practice, June 2008,

Dirik, A.E., Karaküçük, A. 2014. Forensic use of photo response non-uniformity of imaging sensors and a counter method. *Optics Express*, 22(1): 470.

Fechner, N., Kirchner, M. 2014. The humming hum: Background noise as a carrier of ENF artifacts in mobile device audio recordings. Proceedings - 8th International Conference on IT Security Incident Management and IT Forensics, 12-14 May, 2014, Munster, Germany.

Filler, T., Fridrich, J., Goljan, M. 2008. Using Sensor Pattern Noise for Camera Model Identification. In ICIP 2008, Pp. 1296–1299.

Fridrich, J. 2009. Digital image forensics. *IEEE Signal Processing Magazine*, 26(2): 26–37.

Garg, R., Hajj-Ahmad, A., Wu, M. 2013. Geo-location estimation from electrical

network frequency signals. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 26-31 May 2013, 2013, Vancouver, BC, Canada.
Garg, R., Varna, A.L. Hajj-Ahmad, A., Wu, M. 2013. Seeing ENF: Power-signature-based timestamp for digital multimedia via optical sensing and signal processing. *IEEE Transactions on Information Forensics and Security*, 8(9):1417–1432.
Garg, R., Varna, A.L., Wu, M. 2011. Seeing ENF: natural time stamp for digital video via optical sensing and signal processing. Proceedings of the 19th ACM international conference on Multimedia, 28 November-01 December, 2011, Scottsdale, Arizona, USA
Goljan, M., Fridrich, J. 2008. Camera identification from cropped and scaled images. Proc. SPIE 6819, Security, Forensics, Steganography, and Watermarking of Multimedia Contents X, March 2008, San Jose, California, USA.
Gonzalez, R.C., Woods, R.E. 2011. Digital Image Processing. Pearson Education, NJ, USA, 976 pp.
Grigoras, C. 2005. Digital audio recording analysis–the electric network frequency criterion. *Inter- national Journal of Speech Language and the Law*, 12(1):63-76.
Grigoras, C. 2007. Applications of ENF criterion in forensic audio, video, computer and telecom- munication analysis. *Forensic Science International*, 167(2-3):136-145.
Grigoras, C., Cooper, A., Michalek, M. 2009. Best practice guidelines for ENF analysis in forensic authentication of digital evidence. Forensic Speech and Audio Analysis Working Group.
Gu, J., Hitomi, Y., Mitsunaga, T., Nayar, S. 2010. Coded rolling shutter photography: Flexible space-time sampling. IEEE International Conference on Computational Photography, 29-30 March, 2010, Cambridge, MA, USA
Hajj-Ahmad, A., Baudry, S., Chupeau, B., Doërr, G. 2015a. Flicker forensics for pirate device identification. Proceedings of the 3rd ACM Workshop on Information Hiding and Multimedia Security, 17 June, 2015, Portland, Oregon, USA
Hajj-Ahmad, A., Berkovich, A., Wu, M. 2016. Exploiting power signatures for camera forensics. *IEEE Signal Processing Letters*, 23(5):713–717.
Hajj-Ahmad, A., Garg, R., Wu, M. 2012. Instantaneous frequency estimation and localization for enf signals. Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, 3-6 December, 2012, Hollywood, CA, USA
Hajj-Ahmad, A., Garg, R., Wu, M. 2015b. ENF-based region-of-recording identification for media signals. *IEEE Transactions on Information Forensics and Security*, 10(6): 1125-1136
Hua, G., Zhang, Y., Goh, J., Thing, V.L.L. 2016. Audio authentication by exploring the absolute-error-map of ENF signals. *IEEE Transactions on Information Forensics and Security*, 11(5):1003–1016.
Hyun, D.K., Ryu, S.J., Lee, M.J., Lee, J.H., Lee, H.Y., Lee, H.K. 2012. Source camcorder identification from cropped and scaled videos. Proc. SPIE 8303, Media Watermarking, Security, and Forensics, 9 February, 2012, Burlingame, California, USA.
Iuliani, M., Fontani, M., Shullani, D., Piva, A. 2017. Hybrid reference-based Video Source Identification. *Sensors*, 19(3,649): 1-19
Kouokam, E.K., A.E. Dirik, 2019. Prnu-based source device attribution for youtube videos. *Digital Investigation*, 29: 91-100.
Lukáš, J., Fridrich, J., Goljan, M. 2006. Digital camera identification from sensor

pattern noise. *IEEE Transactions on Information Forensics and Security*, 1(2): 205-214.
Mettripun, N., Amornraksa, T., Delp, E.J. 2013. Robust image watermarking based on luminance modification. *Journal of Electronic Imaging*, 22(22): 22-16.
Mondaini, N., Caldelli, R., Piva, A., Barni, M., Cappellini, V. 2007. Detection of malevolent changes in digital video for forensic applications. Proc. SPIE 6505, Security, Steganography, and Watermarking of Multimedia Contents IX, 27 February, 2007, San Jose, CA, USA.
Naumovich, G., Memon, N. 2003. Preventing piracy, reverse engineering, and tampering. *Computer*, 36(7):64–71.
Proakis, J.G., Manolakis, D.G. 2007. Digital Signal Processing: Principles, Algorithms, and Applications. Prentice Hall, 1016 pp.
Savari, M., Wahab, A.W.A., Anuar, N.B. 2016. High-performance combination method of electric network frequency and phase for audio forgery detection in battery-powered devices. *Forensic Science International*, 266: 427–439.
Sencar, H. T., Memon, N. 2013. Digital Image Forensics: There is More to a Picture than Meets the Eye. Springer, New York, NY, US, 372 pp.
Shullani, D., Fontani, M., Iuliani, M., Shaya, O.A., Piva, A. 2017. VISION: a video and image dataset for source identification. *EURASIP Journal on Information Security*, 2017(1):15.
Su, H., Hajj-Ahmad, A., Wong, C.W., Garg, R., Wu, M. 2014a. ENF signal induced by power grid: a new modality for video synchronization. In *Proceedings of the 2Nd ACM International Workshop on Immersive Media Experiences*, 07 November, 2014, Orlando, Florida, USA.
Su, H., Hajj-Ahmad, A., Garg, R., Wu, M. 2014b. Exploiting rolling shutter for ENF signal extraction from video. In 2014 IEEE International Conference on Image Processing (ICIP), 27-30 October 2014, Paris, France.
Su, H., Hajj-Ahmad, A., Wu, M., Oard, D.W. 2014c. Exploring the use of ENF for multimedia synchronization. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 4-9 May, 2014, Florence, Italy.
Tachaphetpiboon, S., Thongkor, K., Amornraksa, T., Delp, E.J. 2014. Digital watermarking for color images in hue-saturation-value color space. *Journal of Electronic Imaging*, 23(3): 1-14.
Taspinar, S., Mohanty, M., Memon, N. 2016. PRNU based source attribution with a collection of seam-carved images. *Pro- ceedings - International Conference on Image Processing, ICIP*, 25-28 September, 2016, Phoenix, AZ, USA.
Vatansever, S., Dirik, A.E. 2016. Forensic analysis of digital audio recordings based on acoustic mains hum. 24th Signal Processing and Communication Application Conference (SIU), 16-19 May, 2016, Zonguldak, Turkey.
Vatansever, S., Dirik, A.E. 2017. The effect of light Source on ENF based video forensics. *Uludag University Journal of The Faculty of Engineering*, 22(1): 53-64.
Vatansever, S., Dirik, A.E., Memon, N. 2017. Detecting the presence of enf signal in digital videos: A superpixel-based approach. *IEEE Signal Processing Letters*, 24(10): 1463–1467.
Vatansever, S., Dirik, A.E., Memon, N. 2019a. Analysis of rolling shutter effect on enf based video forensics. *IEEE Transactions on Information Forensics and Security*, 14(9): 2262 - 2275.
Vatansever, S., Dirik, A.E., Memon, N. 2019b. Factors affecting enf based time-of-recording estimation for video. IEEE International Conference on Acoustics, Speech

and Signal Processing, 12-17 May, Brighon, UK.
Wu, M., Hajj-Ahmad, A., Su, H. 2015. Techniques to extract enf signals from video image sequences exploiting the rolling shutter mechanism; and a new video synchronization approach by matching the enf signals extracted from soundtracks and image sequences. US Patent US9916857B2.

RESUME

Name Surname	: Saffet Vatansever
Place and Date of Birth	: Kardzhali (Bulgaria), 1985
Foreign Languages	: English
Education Status	
High School	: Bursa Boys High School, Turkey - 2003
Bachelor's	: Yildiz Technical University, Turkey - 2007
Master's	: University of Newcastle Upon Tyne, UK - 2013
Work Experience	: Bursa Technical University, Turkey (2014 - …)
Contact (e-mail)	: vatanseversaffet@gmail.com

Publications from the PhD Thesis :

Vatansever, S., Dirik, A.E. 2016. Forensic analysis of digital audio recordings based on acoustic mains hum. 24th Signal Processing and Communication Application Conference (SIU), 16-19 May, 2016, Zonguldak, Turkey.

Vatansever, S., Dirik, A.E. 2017. The effect of light Source on ENF based video forensics. *Uludag University Journal of The Faculty of Engineering*, 22(1): 53-64.

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

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

Vatansever, S., Dirik, A.E., Memon, N. 2019b. Factors affecting enf based time-of-recording estimation for video. IEEE International Conference on Acoustics, Speech and Signal Processing, 12-17 May, Brighon, UK.

Vatansever, S., Dirik, A.E. " Adli Kanıt Kapsamında Analiz Edilen Bir Videoda Elektrik Şebeke Frekansı Sinyalinin Gömülü Olup Olmadığı Bilgisine Ulaşmayı Sağlayan Bir Yöntem, 2016/14117 (in peer review).