ABSTRACT

Title of Document: MICRO SIGNAL EXTRACTION AND ANALYTICS

Chau-Wai Wong, Doctor of Philosophy, 2017

Directed by: Professor Min Wu and Professor Gang Qu
Department of Electrical and Computer Engineering

This dissertation studies the extraction of signals that have smaller magnitudes—typically one order of magnitude or more—than the dominating signals, or the extraction of signals that have a smaller topological scale than what conventional algorithms resolve. We name such a problem the *micro signal* extraction problem.

The micro signal extraction problem is challenging due to the relatively low signal strength. In terms of relative magnitude, the micro signal of interest may very well be considered as one signal within a group of many types of tiny, nuisance signals, such as sensor noise and quantization noise. This group of nuisance signals is usually considered as the "noisy," unwanted component in contrast to the "signal" component dominating the multimedia content. To extract the micro signal that has much smaller magnitude than the dominating signal and simultaneously to protect it from being corrupted by other nuisance signals, one usually has to tackle the problem with extra caution: the modeling assumptions behind a proposed extraction algorithm needs to be closely calibrated with the behavior of the multimedia data. In this dissertation, we tackle three micro signal extraction problems by synergistically applying and adapting signal processing theories and techniques.

In the first part of the dissertation, we use mobile imaging to extract a collection of directions of microscopic surfaces as a unique identifier for authentication and counterfeit detection purposes. This is the first work showing that the 3-D structure at the *microscopic level* can be precisely estimated using techniques related to the photometric stereo. By enabling the mobile imaging paradigm, we have significantly reduced the barriers for extending the counterfeit detection system to end users.

In the second part of the dissertation, we explore the possibility of extracting the Electric Network Frequency (ENF) signal from a single image. This problem is much more challenging compared to its audio and video counterparts, as the duration and the magnitude of the embedded signal are both very small. We investigate and show how the detectability of the ENF signal changes as a function of the magnitude of the embedded ENF signal.

In the last part of the dissertation, we study the problem of heart-rate from fitness exercise videos, which is challenging due to the existence of fitness motions. We show that a highly precise motion compensation scheme is the key to a reliable heart-rate extraction system.

MICRO SIGNAL EXTRACTION AND ANALYTICS

by

Chau-Wai Wong

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2017

Advisory Committee:
Professor Min Wu, Chair/Advisor
Professor Gang Qu, Co-Advisor
Professor Rama Chellappa
Professor K. J. Ray Liu
Professor David W. Jacobs

To my parents and Xinyan.

Acknowledgments

I would like to express my sincere gratitude to my advisor, Prof. Min Wu, for her unwavering guidance throughout the journey of my Ph.D. study. I learned from her not only to have deep insights into a specific problem, but also to gain a high-level understanding beyond the problem. I learned from her to explore challenging research problems by innovatively applying theoretical analyses and developing practical solutions. I also learned from her the importance of and the skills for effectively communicating both in writing and in oral our scientific findings to the research community. These valuable principles will definitely have a long-lasting effect on my future endeavors.

I would like to thank all dissertation committee members, Prof. Gang Qu, Prof. Rama Chellappa, Prof. David Jacobs, and Prof. Ray Liu, for their valuable comments and their excellent courses that laid a solid foundation for my Ph.D. study. I would also like to thank several faculty members from the Math Department, Prof. Paul Smith and Prof. Wojciech Czaja, for their enlightening courses and valuable comments on my research.

I appreciate the mentorship and financial support offered by the Future Faculty Program of the A. James Clark School of Engineering. I would like to thank the National Science Foundation, the ECE Department and the Math Department of the University of Maryland (UMD), for the support during my Ph.D. study. I would also like to thank the Ministry of Education of the Chinese government for selecting me for a prestigious Scholarship Award of Outstanding Self-Financed Students Abroad, which provided recognition, encouragement, and support at the final stage of my Ph.D. study.

I would like to thank Qiang Zhu, Jiahao Su, and Prof. Chang-Hong Fu for their input into the collaborative work on the heart rate extraction project. Dr. Yanpin Ren provided generous help in collecting imaging data for the ENF based geo-tagging project. Dr. King Lam Hui and Zhili Yang kindly offered their valuable expertise in microscopic imaging and taking confocal images for the surface patch work. Part of the surface patch work was in collaboration with AiDiXing, Inc.

It has been my pleasure and privilege to interact with wonderful colleagues at UMD. I would like to thank Wenjun Lu, Wei-Hong Chuang, Wei Guan, Hui Su, Ravi Garg for providing help and guidance in my research and course work. Thanks also go to a number of Media And Security Team (MAST) and Signals and Information Group (SIG) members, in particular, Adi Hajj-Ahmad, Abbas Kazemipour, Qiang Zhu, Yu-Han Yang, Xiaoyu Chu, Hang Ma, Yi Han, Zhung-Han Wu, Xuanyu Cao, Chen Chen, Qinyi Xu, Feng Zhang, Tong Zhou, Wei Li, and Yichen Wang. The intellectual journey without them would be dull, and I enjoyed the discussions with them on a variety of topics.

Last but not least, I would like to thank my parents who support me unconditionally, and my beloved wife and academic partner, Xinyan. I dedicate this dissertation to them.

Table of Contents

1 Introduction 1

1.1 Main Contributions . . . 3

1.1.1 Image Authentication Using Unclonable Feature of Surfaces . 4

1.1.2 Invisible Geo-Location Signature in a Single Image . . . 5

1.1.3 Fitness Heart Rate Measurement Using Face Videos . . . 6

1.2 Dissertation Organization . . . 6

2 Intrinsic Feature of Surfaces for Authentication 8

2.1 Chapter Introduction . . . 8

2.2 Background and Preliminaries . . . 11

2.2.1 Optical Imaging of Paper and Light Reflection Models . . . 11

2.2.2 Paper Patch Registration . . . 14

2.2.3 Authentication via Hypothesis Testing . . . 15

2.3 Paper Authentication Using Scanners and Cameras . . . 16

2.3.1 Norm Maps by Scanners . . . 16

2.3.2 Appearance Images by Cameras . . . 19

2.4 Proposed Image Appearance Based Authentication with Camera Flash 20

2.4.1 Capturing Conditions and Proposed Method . . . 21

2.4.2 Experimental Results . . . 23

2.5 Proposed Surface Norm Based Authentication Using Mobile Cameras 25

2.5.1 Macroscopic Intensity Due to Flashlight . . . 26

2.5.2 Estimating the Normal Vector Field . . . 30

2.5.3 Proposed Method . . . 31

2.5.4 Experimental Conditions and Results . . . 33

2.5.4.1 Totally Dark Environment . . . 34

2.5.4.2 Environment with Ambient Light . . . 35

2.5.4.3 Mobile Cameras of Other Brands . . . 38

2.5.5 Comparison with Prior Work . . . 38

2.6 Perturbation Analysis on Discriminability . . . 39

2.6.1 Precision of Estimated Lens Location . . . 40

2.6.2 Number of Images for Normal Vector Field Estimation . . . 43

2.6.3 Perturbation Factors Combined . . . 45

2.7 Discussions . . . 46

2.7.1 Interpretation of Norm Map From Low Resolution Images . . 46
2.7.2 Effect of Motion Blur 47
2.7.3 Understanding the Physics of Paper Surface Reflection 48
2.7.4 Robustness of the Norm Map 52
2.7.5 Considerations for Using Statistical Methods for Inference . . 53
2.8 Chapter Summary 54

3 High-Contrast, Extrinsic Features of Surfaces for Authentication 56
3.1 Chapter Introduction 56
3.2 Background and Preliminaries 57
3.3 Comparative Study on Using Low-Contrast Feature 58
3.4 Authentication Using High-Contrast Extrinsic Features 62
3.4.1 Lab Produced Low-Cost Sealed Powder Patch 62
3.4.2 Performance Achieved Directly Using Images 63
3.4.3 Performance Achieved Using Spatial Descriptors 63
3.5 Perturbation Analysis 67
3.5.1 Analysis Using Real Images 67
3.5.2 Analysis Using Synthetic Images 70
3.6 Generative Model of High-Contrast Images 72
3.6.1 Decision Statistics as a function of (δ_m, δ_n) 74
3.6.2 Derivation for Joint Distribution $f_{\Delta_m,\Delta_n|H_0}(\delta_m, \delta_n)$ Using Isotropic Foreground Objects 79
3.6.2.1 Matched Case, H_1 79
3.6.2.2 Unmatched Case, H_0 80
3.6.3 Model Validation 85

4 Invisible Geo-Location Signature in a Single Image 87
4.1 Chapter Introduction 87
4.2 Background and Preliminaries 90
4.2.1 Electric Network Frequency 90
4.2.2 Rolling Shutter and Read-Out Time 91
4.2.3 ENF Signal Embedding in a Single Image 92
4.3 Proposed Framework and Approaches 93
4.3.1 Formulation and Framework 93
4.3.2 Independent Component Analysis Based Signal Separation . . 97
4.3.3 Entropy Minimization for Parametric-Surface–Removed Images 99
4.4 Experimental Results 101
4.4.1 Results of Independent Component Analysis Based Method . 101
4.4.2 Results of Entropy Minimization Method 102
4.4.2.1 Images with Synthetic ENF Traces 102
4.4.2.2 Real Rolling Shutter Images with ENF Traces 103
4.5 Discussions 105
4.6 Chapter Summary 106

5 Fitness Heart Rate Measurement Using Face Videos 116

5.1 Chapter Introduction 116

5.2 Proposed Method 118

5.2.1 Precise Face Registration 118

5.2.2 Segment Continuity and Cheek Regions Selection 120

5.2.3 Detrending and Temporal Filtering 122

5.2.4 Motion Frequency Notching 122

5.2.5 Robust Frequency Estimation 123

5.3 Experimental Results 124

5.4 Chapter Summary 128

6 Conclusions and Future Perspectives 129

Bibliography 133

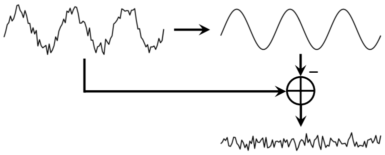
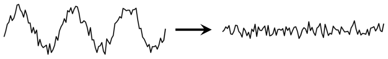
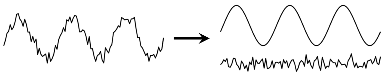
1

Introduction

This dissertation studies the extraction of signals that have smaller magnitudes—typically one order of magnitude or more—than the dominating signals, or the extraction of signals that have a smaller topological scale than what conventional algorithms resolve. We name such a problem the *micro signal* extraction problem.

The micro signal extraction problem is challenging due to the relatively low signal strength. In terms of relative magnitude, the micro signal of interest may very well be considered as one signal within a group of many types of tiny, nuisance signals, such as sensor noise and quantization noise. This group of nuisance signals is usually considered as the "noisy," unwanted component in contrast to the "signal" component dominating the multimedia content. To extract the micro signal that has much smaller magnitude than the dominating signal and simultaneously to protect it from being corrupted by other nuisance signals, one usually has to tackle the problem with extra caution: the modeling assumptions behind a proposed extraction algorithm needs to be closely calibrated with the behavior of the multimedia data.

Table 1.1: Strategies of micro signal extraction for various scenarios

Strategy	**Scenarios**
Residue based 	intrinsic surface feature for authentication [Ch. 2], heart rate from video [Ch. 5], ENF from video [1,2]
Property based 	ENF from image [Ch. 4], ENF from audio [3,4]
Source separation 	ENF from image (limited performance) [Ch. 4]

Directly applying off-the-shelf algorithms would often fail. Instead, a synergistic combination of the ideas behind the algorithms with insightful adaptation may lead to promising results. We summarize in Table 1.1 three strategies toward micro signal extraction, accompanied by scenarios that appeared in literature and in the later parts of the dissertation.

Micro signal extraction is a fundamental enabler for digital/information forensics that is "concerned with determining the authenticity, processing history, and origin of digital multimedia content with no or minimal reliance on side channels other than the digital content itself." [5] In many such applications, forensic investigators heavily depend on the imperceptible or subtle traces buried, namely, the micro signals, in the multimedia content.

1.1 Main Contributions

In this dissertation, we explore micro signal extraction with a focus on the visual aspect using mobile devices. We study two applications in forensics [Ch. 2–4] and one application in noninvasive healthcare monitoring [Ch. 5]. All strategies in Table 1.1 for micro signal extraction are explored.

First, we use mobile imaging to extract a collection of directions of microscopic surfaces as a unique identifier for authentication and counterfeit detection purpose. This is the first work showing that the 3-D structure at the *microscopic level* can be precisely estimated using techniques related to the photometric stereo [6]. By enabling the mobile imaging paradigm, we have significantly reduced the barrier to extending the counterfeit detection system to end users.

Second, we explore the possibility of extracting the Electric Network Frequency (ENF) signal from a single image. This problem is much more challenging compared to its audio counterpart [3, 4] and video counterpart [1, 2], as the duration and the magnitude of the embedded signal are both very small. We investigate and show how the detectability of the ENF signal changes as a function of the magnitude of the embedded ENF signal.

Third, we study the problem of heart-rate from fitness exercise videos, which is challenging due to the existence of fitness motions. We show that a highly precise motion compensation scheme is the key to a reliable heart-rate extraction system.

From the perspective of the *Internet of Things* (IoT) [7–10], this dissertation also contributes to the design of smart nodes/sensors. The IoT has an increasing impact on our everyday life with the technological advancement of smart nodes. Mobile devices with audio-visual sensing capabilities are ubiquitous, and their powerful computing capabilities make them great candidates for smart nodes of the IoT. This dissertation focuses on the extraction of “micro signals” buried in image and video data that effectively converts mobile devices into smart nodes that sense micro signals in the form of physical [Ch. 2], environmental [Ch. 4], and physiological signals [Ch. 5].

Below, we detail the key contributions of this dissertation research.

1.1.1 Image Authentication Using Unclonable Feature of Surfaces

We study the authentication problem of specific pieces of paper by using micro signals extracted from paper surfaces via mobile imaging. Prior work showed high matching accuracy used the *normal vector field*, a unique, microscopic, physically unclonable feature of paper surfaces, estimated by consumer grade scanners. Industrial cameras were also used to capture the appearance of the surface rendered after the normal vector field per the law of optics under a semi-controlled lighting condition. In comparison, past explorations based on mobile cameras were very limited and have not had substantial success in obtaining consistent appearance images due to the uncontrolled nature of the ambient light. We show in this dissertation that images captured by mobile cameras are good for authentication when the camera flashlight is exploited for creating a semi-controlled lighting condition. We propose new algorithms to demonstrate that the normal vector field of a paper surface can be estimated by using multiple camera-captured images of different viewpoints. Perturbation analysis shows that the proposed method is robust to inaccurate estimates of camera locations, and a matching accuracy of 10^{-4} in equal error rate (EER) can be achieved using 6 to 8 images under a lab-controlled ambient light environment. Our findings can relax the restricted imaging setups to enable paper authentication under a more casual, ubiquitous setting with a mobile imaging device, which may facilitate duplicate detection of paper documents and counterfeit mitigation of merchandise packaging.

1.1.2 Invisible Geo-Location Signature in a Single Image

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

1.1.3 Fitness Heart Rate Measurement Using Face Videos

Recent studies showed that subtle changes in human's face color due to their heartbeat can be captured by digital video recorders. Most work on extracting such micro signals focused on still/rest cases or those with relatively small motions. In this dissertation, we propose a heart-rate monitoring method for fitness exercise videos. We focus on building a highly precise motion compensation scheme with the help of the optical flow, and use motion information as a cue to adaptively remove ambiguous frequency components for improving the heart rate estimates. Experimental results show that our proposed method can achieve highly precise estimation with an average error of 1.1 beats per minute (BPM) or 0.58% in relative error.

1.2 Dissertation Organization

The rest of the dissertation is organized as follows. In Chapter 2, we first examine the authentication performances using the intrinsic feature of paper surfaces when restricted imaging setups are used, which serve as a performance baseline. We propose methods working under a more flexible setup—mobile cameras with built-in flashlights, and compare the performances with prior work. We conduct perturbation analysis to demonstrate the practicality of the proposed mobile cameras-based authentication method. We then use confocal microscopic data from a physics aspect to elucidate a deeper understanding of the proposed work, and also address some practical issues. In Chapter 3, we study a related authentication problem using extrinsic features of surfaces.

In Chapter 4, we first review the background information on the ENF signature and the image capturing operations that allow it to be present in images of interest. We then present our proposed approaches for extracting the ENF signature from a single image. We show the experiments we conducted and the results obtained.

In Chapter 5, we propose our video-based heart-rate monitoring method specially designed for fitness exercises. We present the experimental results with comparisons if some modules were otherwise replaced or turned off.

Finally, in Chapter 6, we conclude this dissertation and outline research issues for future explorations.

2

Intrinsic Feature of Surfaces for Authentication

2.1 Chapter Introduction

Merchandise packaging and valuable documents such as tickets and IDs are common targets for counterfeiting. Low-cost surface structures have been exploited for counterfeit detection by using their optical features. The randomness of the surface makes the structures physically unclonable or difficult to clone to deter duplications. Such surface structures can be extrinsic by adding ingredients such as fiber [11,12], small plastic dots [11], air bubble [11], powders/glitters [13] that are foreign to the surface; and the surface structures can also be intrinsic by exploring the optical effect of the microscopic roughness of the surface, such as the paper surface formed by inter-twisted wood fibers [13–17]. The inherent randomness of the microscopic roughness quantified using the *normal vector field* has been used as a feature for the unique identification of a particular patch of a surface in [13, 14].

In this chapter, we focus on the intrinsic property of the paper surface for counterfeit detection and deterrence, and seek to find a more casual, ubiquitous imaging setup using consumer grade mobile cameras under commonly available lighting conditions. The previous work in [13–15] shows that the microscopic roughness of the paper surface can be optically captured by consumer grade scanners and industrial cameras, both under controlled lighting conditions in the form of image appearance rendered according to the physical law of light reflection at the paper surface. The appearance images, and the subsequent normal vector field of the surface estimated from the appearance images, can achieve satisfactory authentication results. However, recent work in [13, 17] also showed that if the ambient lighting is not well controlled, the image appearance alone has not achieved satisfactory authentication results. Instead, features based on intensity gradient of visually observable dots are less sensitive to the change of lighting and may be used for authentication at the cost of higher algorithm complexity and moderate discrimination capabilities [17].

Satisfying two requirements may facilitate paper authentication via mobile cameras. First, the mobile captured images should be comparable in resolution and contrast to those captured by scanners. Second, lighting should be controlled to render a desirable appearance of the paper. The first requirement can be qualitatively checked by comparing the acquired images from scanners and mobile cameras. Images acquired in both ways do have similar, detailed intensity fluctuations when zoomed in. The second requirement can be fulfilled by activating the flashlight next to the camera lens on mobile devices. The desirable appearance of the surface can be reasonably expected from the geometric arrangement between the camera and the surface.

As we shall show in this chapter, camera flashlights exploited for creating semi-controlled lighting conditions can significantly improve the performance of using appearance images as the authentication feature. More importantly, by exploiting the underlying rendering principle of the appearance of the surface, *i.e.*, the fully diffuse reflection model [14, 18], one can estimate the normal vector field of the surface without resorting to more restricted acquisition conditions. To the best of our knowledge, this work together with its preliminary version [19] is the first set of work using mobile cameras to obtain an effective estimate of the normal vector field of the paper surface for authentication. In this journal version, additional experimental results are presented for more practical capturing scenarios. Extended perturbation analyses of discrimination power on two factors, namely, the inaccuracy of estimated camera locations and the number of images used for normal vector field estimation, are conducted and explained using statistical methods. "Ground-truth" 3-D structure of paper surface is obtained with confocal microscopy in order to quantitatively examine the linkage between the appearance and the physical structure of the paper surface.

This chapter is organized as follows. In Chapter 2.2, we review light reflection models, the method for paper surface registration, and the method for paper authentication. In Chapter 2.3, we examine the authentication performances when restricted imaging setups are used, which serve as a performance baseline. In Chapters 2.4 and 2.5, we propose methods working under the a more flexible setup—

mobile cameras with built-in flashlights, and compare the performances with prior work. In Chapter 2.6, we conduct perturbation analysis to demonstrate the practicality of the proposed mobile cameras based authentication method. In Chapter 2.7, we use confocal microscopic data from a physics aspect to elucidate a deeper understanding in the proposed work, and also address some practical issues. In Chapter 2.8, we conclude the chapter.

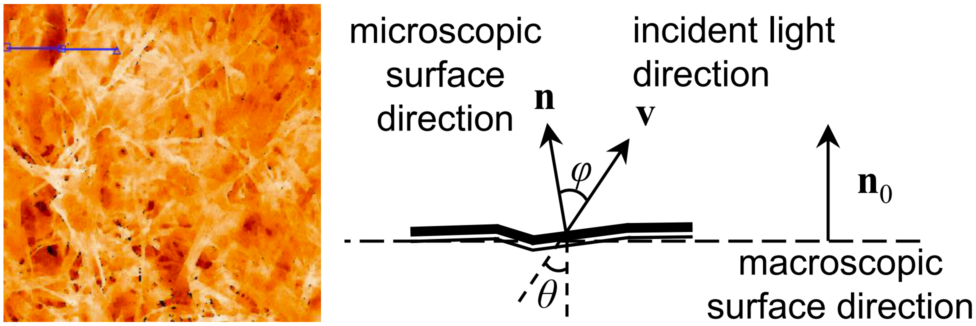
2.2 Background and Preliminaries

2.2.1 Optical Imaging of Paper and Light Reflection Models

Seemingly smooth paper surfaces contain inherent microscopic 3-D structure due to overlapped and inter-twisted wood fibers. This microscopic structure is different from one paper to another and even one location to another on the same paper, and therefore can serve as a unique identifier or fingerprint. One quantitative feature of such 3-D structure, the surface direction, has been successfully exploited for authentication in [13–15].

Fig. 2.1(a) shows a topographic map of a 1 mm-by-1 mm region of a paper surface estimated from images captured by a confocal microscope [20]. The microscopic roughness due to fibers is clearly shown. The visual appearance of the surface follows the law of optics.

Geometrical light reflection models such as specular model and diffuse model have been widely used in computer vision/graphics applications, due to their good approximations to the law of optics and relatively simple analytical forms [6,18,21].



(a) (b)

Figure 2.1: (a) A topographic map of a 1mm-by-1mm region of a paper surface captured by a confocal microscope, reproduced from [20]. The pseudo-color represents the elevation of fibers in the z-direction. (b) Microscopic view of a particular spot on paper surface. Note that φ and θ are not co-planar in most cases. All vectors are unit vectors.

Under the *specular* reflection model, the perceived intensity is dependent on the direction of the reflected light and the direction of eye/sensor. Under the *diffuse* reflection model, the perceived intensity is dependent on the direction of incident light and normal direction of the microscopic surface. The appearances of most surfaces contain both reflection components.

Previous authentication work [13, 14] treating paper as a fully diffuse surface has led to satisfactory results. We follow fully diffuse modeling assumption, and provide an experimental justification in the discussion section that the strengths of diffuse component versus the specular component is about six to one.

Fig. 2.1(b) shows the microscopic surface normal direction, $\mathbf{n}$, of a particular spot $\mathbf{p}$ in a microscopic view (which is often different from the macroscopic surface direction, $\mathbf{n}_0$), and an incident light direction, $\mathbf{v}$. The perceived reflected intensity l_r of the fully diffuse reflectance model [14,18] is

$$l_r(\mathbf{p}) = \lambda \cdot l(\mathbf{p}) \cdot \underbrace{\mathbf{n}(\mathbf{p})^T \mathbf{v}(\mathbf{p})}_{=\cos\varphi(\mathbf{p})}, \tag{2.1}$$

which depends on the angle $\varphi(\mathbf{p})$ between normal direction of the surface at the microscopic level, $\mathbf{n} = (n_x, n_y, n_z)$, and the direction where the incident light is coming from, $\mathbf{v} = (v_x, v_y, v_z)$; the strength of the light at the current spot, $l(\mathbf{p})$; and the albedo, λ, characterizing the physical capability of reflecting the light [6,21]. In our work, the assumption of λ being constant over the whole paper patch is found to hold well for the purpose of authentication.

For an ideal point light source, the light strength $l(\mathbf{p})$ over a spatial field is modeled by considering the effect of energy fall-off due to travel distance of the light [6]. In practice, a camera flashlight is not a perfect point source but has a finite dimension, *e.g.*, in a disc-like shape. When the flashlight is not perfectly oriented towards the paper surface, it can lead to a foreshortening effect reducing the strength of the light arriving at the projected point of the light on paper. Therefore, it is practically difficult to model $l(\mathbf{p})$ with a high precision for nonideal point sources. Instead, we estimate $l(\mathbf{p})$ by exploiting its spatial smoothness property. With the values of $l(\mathbf{p})$, the microscopic structure can then be determined in terms of normal vectors, $\mathbf{n}(\mathbf{p})$.

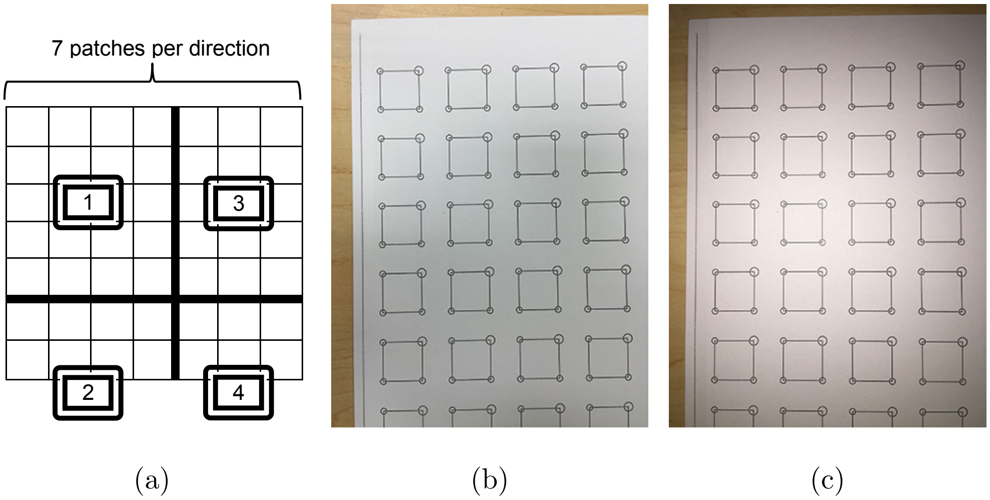


Figure 2.2: (a) Four camera shots are needed for capturing 49 square patches located on a piece of paper. (b) Image captured at position#1 under ambient (fluorescent) light without flash (database 505), and (c) with flash (database 501). Capturing device: iPhone 6.

2.2.2 Paper Patch Registration

We use a simple square-shaped registration container from our recent work [13] as shown in Fig. 2.2(b), and a tri-patch extension as shown in Fig. 2.6(b), to facilitate the precise registration in our experiments. Considering a printing resolution of 600 pixels per inch, each square container of size $\frac{2}{3}$-by-$\frac{2}{3}$ inch^2 (1.69-by-1.69 cm^2) corresponds to a box of 400-by-400 in pixel at a line width of 5 pixels, and there are four circles at corners of each square. A preliminary alignment based on four boundaries can be achieved using a Hough transform, and subpixel resolution refinement with perspective transform compensation is then carried out based on the circle markers. Lens location relative to the surface in the world coordinate system can be readily calculated from the estimated perspective transform matrix, and then the direction of incident light at every pixel location is known. Note that the world coordinate system is naturally defined to have the xy-plane located at the bottom plane of the paper surface and the z-axis pointed upwards. All camera captured images were unwarped to remove the effect of lens distortion before being used if they were captured by camera. This step improves the matching performance on average by 0.04 in terms of correlation value in our experiments.

2.2.3 Authentication via Hypothesis Testing

We approach the patch verification problem as a binary hypothesis testing problem [22] using discriminative features derived from images of the paper. The null hypothesis H_0 is that the test/query patch does not match with the patch from the reference database, whereas the alternative hypothesis H_1 is that the test/query patch matches with the reference patch. To quantify the degree of match, we use the normalized sample correlation $\hat{\rho}$ on the pair of extracted features, *e.g.*, pixel intensity in Chapter 2.4 and surface normal vector in Chapter 2.5. We estimate the probability density functions (PDFs) $f_{\hat{\rho}|H_0}(\hat{\rho})$ and $f_{\hat{\rho}|H_1}(\hat{\rho})$ that have very distinct mean values under unmatched and matched cases, and make a decision using the simple thresholding rule on an observed value of the random variable $\hat{\rho}$.

Under the simple thresholding rule with threshold τ, the detection rate is defined to be $P_D(\tau) = \int_{\tau}^{\infty} f_{\hat{\rho}|H_1}(\xi)\, d\xi$ [or its complement, the miss rate, $P_M(\tau) = 1 - P_D(\tau)$] and the false-alarm rate is defined to be $P_F(\tau) = \int_{\tau}^{\infty} f_{\hat{\rho}|H_0}(\xi)\, d\xi$. The receiver operating characteristic (ROC) curve $\big(P_F(\tau), P_D(\tau)\big)$ can be drawn by varying the value of τ to reveal the discrimination capability of the system. Alternatively, the equal error rate (EER), $\{P_{\text{EE}} \,|\, P_{\text{EE}} = P_F(\tau) = P_M(\tau), \tau \in \mathbb{R}\}$, can be used as a compact, one-score indicator for the discrimination capability. For Gaussian and Laplacian distributions, it is not difficult to derive the analytical forms of EER to be $\Phi\left(\frac{\mu_0-\mu_1}{\sigma_0+\sigma_1}\right)$ and $\frac{1}{2}\exp\left[\frac{\lambda_0\lambda_1}{\lambda_0+\lambda_1}(\mu_0-\mu_1)\right]$, respectively [22,23], where $\Phi(\cdot)$ is the cumulative density function for the standard Gaussian distribution. The theoretical quantities, including the mean μ_i, standard deviation σ_i, and rate λ_i, can be replaced by their estimates from the real data. In this way, EER can be estimated even when there are a relatively limited number of data points and/or the PDFs are widely separated in which the true tails of the PDFs may not be adequately revealed by simulated data. More discussions on using practical data for the theoretical model described above can be found in Chapter 2.7.5.

2.3 Paper Authentication Using Scanners and Cameras

2.3.1 Norm Maps by Scanners

The norm map as a physical feature of a paper surface has been found to have strong discrimination power. Clarkson *et al.* [14] used the fully diffuse reflectance model as described in Eq. (2.1) to estimate the projected normal directions at all integer-pixel locations of the surface. We refer to the collection of normal vectors for all pixels as the *normal vector field* (containing x-, y-, and z-components), and its projection onto surface plane as the *norm map* (containing x- and y-components only). A norm map can be estimated using images scanned from four different orientations of the paper: 0°, 90°, 180°, and 270°. Without knowing the exact direction of an incident light, an estimate of one component of the norm map can be obtained as the difference between two scans in exactly opposite directions, canceling the effect of the unknown incident direction of the scanner light. The norm map containing randomly distributed vectors has been used as a feature for the unique identification of a particular patch of a surface in [13, 14]. In [14], a seeded hash is computed by random projection, and the Hamming distance of two hashes is used as the decision statistic. The sample statistics such as mean and variance measured from Fig. 8 of [14] reveal the EER to be between 10^{-130} and 10^{-15} (see Table 2.6 for comparison) per our discussion on estimating the EER in Chapter 2.2.3.

In order to provide more accurate norm map estimates as the reference data for our proposed method in Chapter 2.5, we improve the norm map estimation algorithm over those in [13, 14] by removing the global bias for x- and y-components of the estimated norm map. Below we carry out experiments using the improved norm map estimator to provide a baseline for comparisons in later sections.

We estimated norm maps for 49 distinct square-shaped patches located on a piece of paper. The acquisition procedure was repeated using two Epson scanners: Perfection 2450 and GT-2500. Sample patches for scanner 2450 and the resulting norm map estimate of 1/100 of the patch size are shown in Fig. 2.3. Authentication using the hypothesis testing described in Chapter 2.2.3 was carried out by correlating the test feature with the reference feature. Three features, namely, the normal vector's length, x- and y-components, were tested and the results are shown

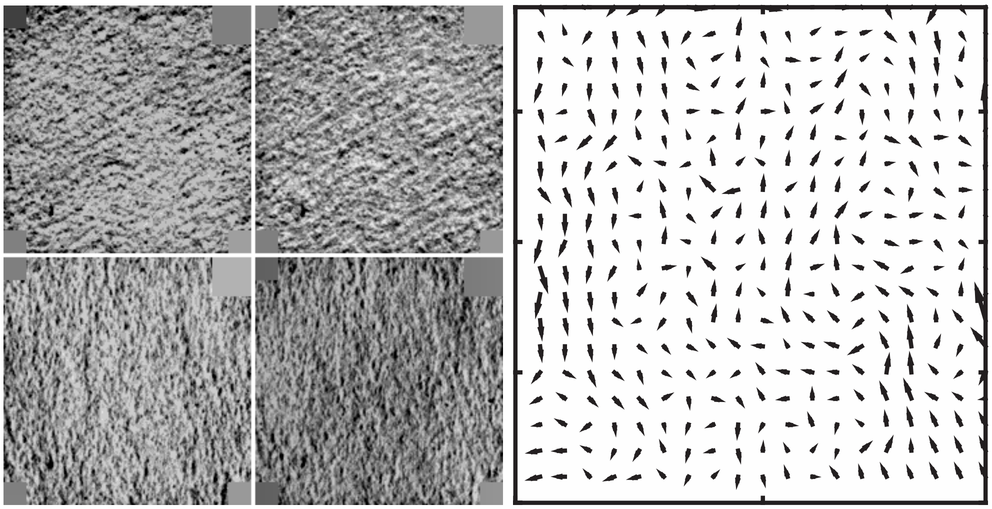


Figure 2.3: Scanned images from four perpendicular orientations of a piece of $\frac{2}{3}$-by-$\frac{2}{3}$ inch^2 (1.69-by-1.69 cm^2) paper, and the resulting estimated norm map covering 1/100 area of that paper.

in the three columns of Fig. 2.4, respectively. Each plot contains two estimated PDFs of sample correlation coefficient $\hat{\rho}$: one for matched cases (H_1), and the other for unmatched cases (H_0). All six plots reveal that the distributions for the two hypotheses are far apart and have no overlap, thus a threshold can be set to have no false alarm and no miss detection, suggesting an excellent authentication performance. In addition, the performance for the intra-scanner case (*i.e.*, both test and reference data were obtained using the same scanner) shown in the first row of Fig. 2.4 is slightly better than that for the inter-scanner case (*i.e.*, test and reference data were obtained using different scanners) shown in the second row. They reveal that different acquisition devices can give slightly inconsistent norm map estimates but the inconsistency is not strong.

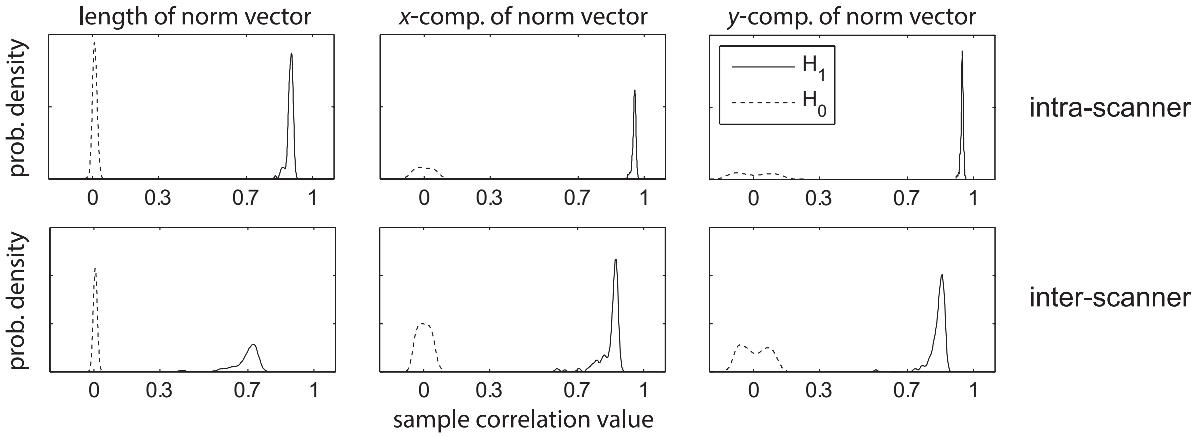


Figure 2.4: Estimated PDFs of sample correlation coefficient $\hat{\rho}$ for unmatched (H_0) and matched (H_1) cases. First row (intra-scanner): Datasets #2–3 (test) vs. #1 (ref.) of scanner 2450, and second row (inter-scanner): Datasets #1–3 (test) of scanner GT vs. #1 (ref.) of scanner 2450. Features: length (column 1), x-component of normal vector (column 2), and y-component of normal vector (column 3).

2.3.2 Appearance Images by Cameras

Instead of using scanners to capture images with directional linear light and closely placed imaging sensors, Voloshynovskiy *et al.* [15,16] examined the imaging setup of using two industrial cameras with a semi-controlled lighting condition—a fixed, ring-shaped light source. The resulting images have similar appearances during multiple capturing instances due to the semi-controlled lighting conditions. The ROC curves from [15] reveal that the EER is around 10^{-4}.

Mobile cameras were used to test the authentication performance under uncontrolled ambient light in [13]. The uncontrolled light can lead to unpredictable surface appearances per the light reflection model in Eq. (2.1). Even using newer mobile cameras such as the iPhone 6 with improved acquisition quality over the older mobile devices, the authentication performances under uncontrolled ambient light are still limited, as revealed by Fig. 3 of [13]. One way to improve the authentication performance as shown by Diephuis *et al.* [17] is to use the intensity gradient based features, *e.g.*, scale-invariant feature transform (SIFT), of high contrast spots that are less sensitive to the change of lighting, at the cost of increasing the design complexity of the authentication system. Extrapolating data points from Table 2 of [17] into the ROC plot of Fig. 8 of [17], we estimate the performance of the proposed SIFT-based method to be around 10^{-2} in terms of EER (see Table 2.6 for comparison).

2.4 Proposed Image Appearance Based Authentication with Camera Flash

Inspired by the success of the approaches discussed in Chapter 2.3 in which lightings for image acquisition are well controlled, we explore a semi-controlled lighting condition with the help of the built-in flashlight of mobile cameras for authentication. We achieve the semi-controlled lighting condition by exploiting the fact that relative positions among the light source, the lens, and the paper patch are known or can be estimated from a captured image.

The simplest case, presented in Chapter 2.4 (this section), is to use the appearance of the patches when cameras are positioned at the same location relative to

Table 2.1: Capturing Conditions for Various Databases

Database ID	502	503	506	508	509	510	501	511	505
Lighting	flash only	flash + ambient light							ambient light only
Device	iPhone 6			iPhone 5s		iPhone 5	iPhone 6	Canon SX230HS	iPhone 6
Room Size	small		large	small	large	small			large

the physical patch so that the effect of lighting is the same for instances of capturing test and reference images. A more sophisticated case, presented in Chapter 2.5 (the next section), is to exploit the physics of lighting and to use multiple images for estimating the normal vector field as the feature for authentication.

2.4.1 Capturing Conditions and Proposed Method

Patches were acquired by the built-in cameras of three mobile devices, iPhone 6, iPhone 5s, and iPhone 5, with and without a flash. The capturing process was done in a large room with 12 overhead fluorescent light arrays, and in a small room with 2 overhead fluorescent light arrays, respectively. The device were held by hand approximately in parallel with the surface of a piece of paper and at a height about $z = 15.5\,\text{cm}$. Detailed *capturing conditions* and the corresponding database ID that will be referred to in the remaining part of this section are shown in Table 2.1.

We use a total of 49 distinct square-shaped paper patches for the experiment. To acquire a *database* of a particular capturing condition, we captured three images for every patch, with slight camera rotation and panning among different capturing instances. Within each database, we refer to the 49 patches for the ith capturing

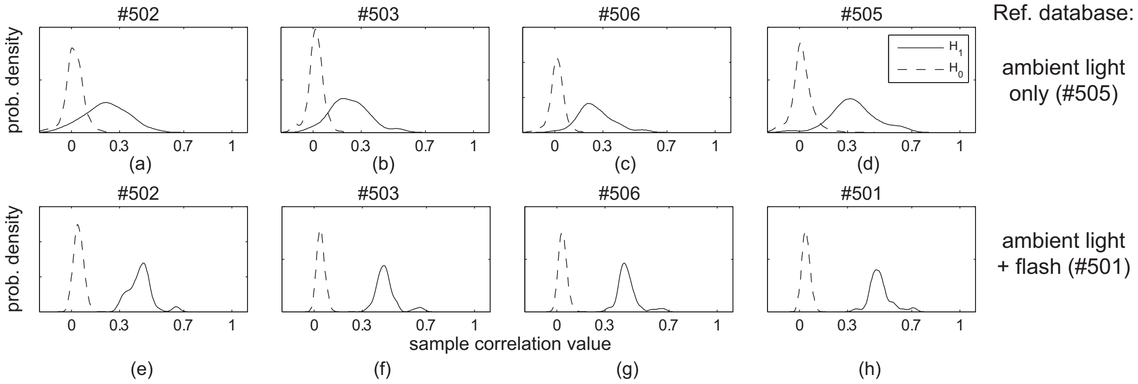


Figure 2.5: First row: authentication performances of 4 test databases vs. reference database #505 (ambient light only). Second row: performances of 4 test databases vs. reference database #501 (flash + ambient light, proposed). Capturing device: iPhone 6.

instance as Dataset $\#i$, for $i = 1, 2, 3$. To speed up the capturing process, patches were acquired together with neighboring patches located on the same piece of paper. A total of four shots were needed to capture the whole region containing the patches, and the camera positions relative to the paper are shown on Fig. 2.2(a). Boundaries among different shots are separated by the thick lines. Figs. 2.2(b) and (c) containing luminance non-uniformity were acquired without and with flashlight for the top-left 20 patches on the layout of the paper.

The way of capturing the whole database of patches as laid out above ensures that any pair of matched test and reference images are captured at the same location relative to the physical patch. That is, the incident light for the test and reference images are the same, effectively controlling the acquisition conditions. In this way, the perceived intensities of patches are similar across different capturing instances per the fully diffuse light reflection model in Eq. (2.1). We shall examine in Chapter 2.5 how an authentication scheme should be designed when we do not constrain the relative locations of the cameras to the physical patches.

Each patch in the captured image was extracted, warped, and registered to a grid of 200-by-200 pixels using the registration procedure outlined in Chapter 2.2.2. In this experiment, images captured at the height of about $z = 15.5\,\text{cm}$ contain around 300 pixels along the edge/side for each patch in raw images, which are of enough resolution (1.25× more pixels than necessary) to generate the registered images. The collection of the 200×200 pixel values of the registered image is then considered as a feature for the paper patch, and normalized sample correlation $\hat{\rho}$ between features from test and reference patches can be calculated for authentication using the hypothesis testing described in Chapter 2.2.3. In this experiment, Datasets #2 and #3 of a database is considered as the test data, and Dataset #1 of the same or a different database is considered as the reference data.

2.4.2 Experimental Results

The contrast of the PDF plots between the first row and second row in Fig. 2.5 shows a significant improvement due to the use of the camera flash. The plots in the first row of Fig. 2.5 reveal that when the test patches with a flash are matched against reference patches without a flash, the authentication performances are limited. A representative plot such as Fig. 2.5(b) has an EER of around 10^{-1}. The plots in the second row of Fig. 2.5 reveal that when both test and reference patches are captured under camera flash as we proposed, the authentication performances are

Table 2.2: Modes of estimated PDFs of correlation[1] for matched cases (H_1). Contrasting conditions: using flash or not, and small vs. large room size.

Ref / Test	With Camera's Built-in Flash				No flash
	Small Room			Large Room	
	501	**502**	**503**	**506**	**505**
501	*0.48*	*0.44*	*0.43*	*0.40*	0.21
502	*0.45*	*0.45*	*0.47*	*0.40*	0.22
503	*0.43*	*0.47*	*0.48*	*0.41*	0.19
506	*0.42*	*0.45*	*0.42*	*0.43*	0.21
507	*0.40*	*0.41*	*0.40*	*0.40*	0.23
505	0.21	0.24	0.24	0.24	0.32

[1] The italic numbers in this table and Table 2.3 correspond to the scenarios that estimated PDFs for matched (H_1) and unmatched (H_0) cases can be perfectly separated.

good and the ambient lighting conditions do not have a major negative effect on the performances. A representative plot such as Fig. 2.5(f) has an EER of around 10^{-5} to 10^{-3} (see Table 2.6 for comparison) per our discussion on estimating the EER with different probabilistic models in Chapter 2.2.3.

Tables 2.2–2.3 present the comprehensive results of various combinations of test and reference databases. Table 2.2 reveals that the flash is the dominating factor to the authentication performance whereas the condition of the ambient light is not an important factor. Table 2.3 reveals that good authentication performance can be achieved across devices of similar imaging modules. The slightly lowered performances for the combinations of iPhone and Canon cameras can be attributed to the different imaging configurations for the two brands of cameras, such as the pattern of the flashlight, and the relative position of the flash module to the lens.

Table 2.3: Modes of estimated PDFs of correlation under for matched cases (H_1). Contrasting condition: camera model.

Ref Test	iPhone 6		iPhone 5s		iPhone 5	Canon SX230
	503	**506**	**508**	**509**	**510**	**511**
503	*0.48*	*0.41*	*0.47*	*0.44*	*0.36*	0.31
506	*0.42*	*0.43*	*0.44*	*0.42*	*0.36*	0.29
508	*0.46*	*0.42*	*0.52*	*0.47*	*0.38*	0.33
509	*0.44*	*0.41*	*0.47*	*0.50*	*0.37*	0.32
510	*0.37*	*0.35*	*0.38*	*0.37*	*0.35*	0.24
511	0.28	0.26	0.32	0.30	0.23	0.26

2.5 Proposed Surface Norm Based Authentication Using Mobile Cameras

Although the authentication scheme using flashlight proposed in Chapter 2.4 outperforms schemes that flashlight are not used, the requirement that the test and reference images must be captured at the same position relative to the physical patch is not practical. As mobile cameras have ever increasing acquisition quality, we ask a research question: Is it possible to use mobile cameras to estimate the physical feature of the paper—the normal vector field—by using multiple images, while solving the issues of camera geometry and lighting? Photometric stereo approaches have long been used to reconstruct 3-D surfaces using photos of surfaces [6]. However, the challenge here is that the scale of interested surface in our problem here is much smaller. We therefore need to carefully examine the physics of light reflection and arrive at a light reflection model with a proper level of sophistication, in order to obtain meaningful estimates of the normal vector field.

2.5.1 Macroscopic Intensity Due to Flashlight

Examining the images captured under the flash in Fig. 2.2 (c), one can observe that there exists a mild spatial intensity change across the image. Examining the reflective intensity of a small region under a strong light both by human eyes and from the digital image, one can observe a high spatial frequency fluctuation in addition to the mild intensity change. Given that the light intensity arriving at the paper slowly varies spatially, this fluctuation of the reflective intensity is therefore attributed mainly to the inconsistent orientations of the paper surface at the microscopic level. To reveal the intensity change in fine details, the mild change at the macroscopic level should be first removed. We define the image intensity at the macroscopic level as the macroscopic intensity, l^{macro}.

It can be shown as follows that the macroscopic intensity l^{macro} is proportional to the light strength arriving at the surface, l, and cosine of the incident angle, θ. We approximate the macroscopic intensity by the *averaged perceived intensity* $\overline{l_r}$ of background pixels over a small neighborhood $\mathcal{N}$ around a pixel location $\mathbf{p}$:

$$\begin{aligned}
l^{macro}(\mathbf{p}) &\approx \overline{l_r}(\mathbf{p}) && \text{(2.2a)}\\
&= \frac{1}{|\mathcal{N}(\mathbf{p})|} \sum_{\mathbf{k} \in \mathcal{N}(\mathbf{p})} \lambda \cdot l(\mathbf{k}) \cdot \mathbf{n}(\mathbf{k})^T \mathbf{v}(\mathbf{k}) && \text{(2.2b)}\\
&\overset{(a)}{\approx} \lambda \cdot l(\mathbf{p}) \cdot \left[\tfrac{1}{|\mathcal{N}(\mathbf{p})|} \sum \mathbf{n}(\mathbf{k}) \right]^T \mathbf{v}(\mathbf{p}) && \text{(2.2c)}\\
&\overset{(b)}{\approx} \lambda \cdot l(\mathbf{p}) \cdot \mathbb{E}\left[\mathbf{n}(\mathbf{p})\right]^T \mathbf{v}(\mathbf{p}) && \text{(2.2d)}\\
&\overset{(c)}{=} \lambda \cdot l(\mathbf{p}) \cdot [0, 0, \mu_{n_z}]\, \mathbf{v}(\mathbf{p}) && \text{(2.2e)}\\
&= \lambda \cdot l(\mathbf{p}) \cdot \mu_{n_z} \cdot \underbrace{v_z(\mathbf{p})}_{\cos\theta \text{ at } \mathbf{p}} && \text{(2.2f)}
\end{aligned}$$

where $|\mathcal{N}(\mathbf{p})|$ is the number of pixels in the small neighborhood of $\mathbf{p}$; step-a follows from the fact that $l(\mathbf{k})$ and $\mathbf{v}(\mathbf{k})$ are approximately constant over the small neighborhood; step-b follows from ergodicity; and step-c follows from the assumption that the normal vectors in the world coordinate system defined in Chapter 2.2.2 are on average pointing straight up, *i.e.*, $\mathbb{E}[n_x] = \mathbb{E}[n_y] = 0$ and $\mathbb{E}[n_z] = \mu_{n_z}$, where μ_{n_z} is a modeling constant between 0 and 1.

The smooth nature of the macroscopic intensity l^{macro} over the spatial coordinates makes parametric surfaces promising candidate estimators. In this work, we fit a high-order polynomial surface directly to an image captured under flashlight using an iteratively reweighted least-squares method. The bisquare weights [24] were used to gradually lower the impact of outliers as iteration goes on. The original image and its parametrically fitted version are shown in Figs. 2.6 (b) and (c). As our objective is to obtain the macroscopic intensity due to the flashlight, the image pixels belonging to the registration container and the QR code are considered to be outliers for the surface fitting purpose. The fitting was excellent with almost no bias. The sample standard deviation, about 2 out of 256 shades of gray, quantifies the magnitude of the fine details of the image appearance of the paper. Fig. 2.6 (d) shows a representative row of pixels (with outliers) and its fitted curve.

One should note that even though a detrended patch image can be obtained by pixel-wise division of macroscopic intensity l^{macro}, the detrended patch image is not suitable directly for authentication via correlating with some reference image. After detrending, images for the same patch captured with flashlight/camera located at different relative locations to the physical patch can have different visual appearances at a small scale. This is caused by the different incident light directions with respect to the microscopic surfaces. Fig. 2.6(e) shows four such detrended patch images when camera locations were at the four corners to the patch. They appear similar at a large scale after detrending, but are very different at a small scale due to different incident light directions.

Fig. 2.7 shows the averaged correlations among the detrended patches as a function of the horizontal and vertical differences in patch locations $(\Delta N_x^{\mathrm{p}}, \Delta N_y^{\mathrm{p}})$, or equivalent in camera capturing locations $(\Delta N_x, \Delta N_y)$. The figure reveals that the farther the capturing distances of cameras for two patches is, the lower the correlation can be for the detrended patch images. This is reasonable as more change in the direction of the incident light leads to more change in microscopic appearance of the paper surface. This implies that without the proper constraints of the relative position between the camera and the patch, it may not be sensible to verify a paper surface using its detrended image, as the correlation value can be unpredictable and

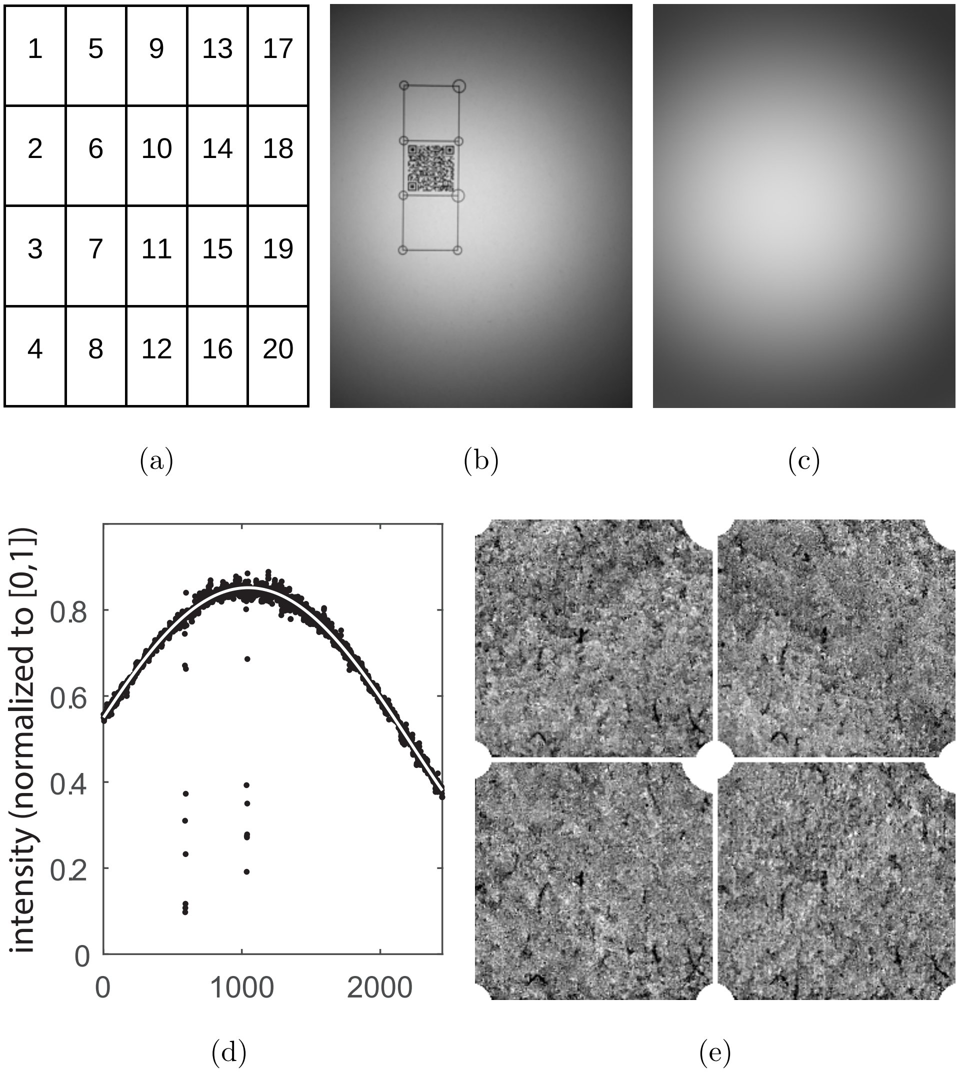


Figure 2.6: (a) A total of $M = 20$ indexed locations for captured patches in images, (b) a paper patch at location #6 of Session 6, paper #920, (c) its estimated macroscopic intensity image l^{macro} obtained by fitting an order-$(5, 5)$ polynomial surface, (d) a row of intensity values from the middle of the image in (b) and the fitted curve, and (e) detrended images (with contrast enhancement) of Session 6 for the paper patch #920 at locations #1, #17, #20, and #4, respectively, shown clockwise.

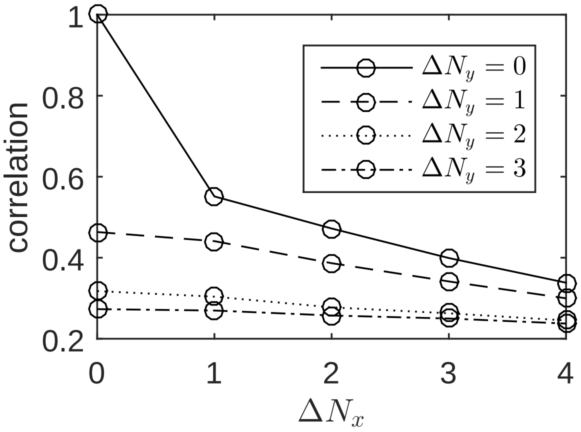


Figure 2.7: Averaged correlation values of detrended images as a function of the distance between camera capturing locations, namely, $(\Delta N_x, \Delta N_y)$.

not a single threshold can be selected to determine the authenticity. This observation further justifies the need of using normal vectors for authentication instead of using images directly.

2.5.2 Estimating the Normal Vector Field

In order to solve for the normal vectors $\mathbf{n}(\mathbf{p})$, we combine Eq. (2.1) characterizing the pixel-wise intensity and Eq. (4.1) characterizing the macroscopic intensity via canceling their common term $\lambda \cdot l(\mathbf{p})$. One can arrive at the following equality by grouping constants and known terms to the left-hand side:

$$\zeta(\mathbf{p}) \approx \mathbf{n}(\mathbf{p})^T \mathbf{v}(\mathbf{p}) \tag{2.3}$$

where $\zeta(\mathbf{p}) = \mu_{n_z} v_z(\mathbf{p}) \cdot l_r(\mathbf{p}) \,/\, l^{\,macro}(\mathbf{p})$ is defined as the normalized intensity and contains the unknown modeling constant μ_{n_z}, the image acquired under flashlight l_r, and the already estimated terms $l^{\,macro}$ and v_z. On the right-hand side, normal vectors $\mathbf{n}(\mathbf{p})$ is yet to be solved, and incident light direction $\mathbf{v}(\mathbf{p})$ is known from previous estimation. The inference problem of normal vector field can therefore be restructured into a linear regression problem when Eq. (2.3) is overdetermined.

More specifically, we estimate the normal vectors independently at every pixel location for a total of 200×200 pixels. For each pixel location $\mathbf{p}$, we setup a system of linear equations using $M = 20$ acquired images, where M is far greater than 4, the number of unknowns:

$$\underbrace{\begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \vdots \\ \zeta_M \end{bmatrix}}_{\zeta} = \underbrace{\begin{bmatrix} \mathbf{v}_1^T & 1 \\ \mathbf{v}_2^T & 1 \\ \vdots & \vdots \\ \mathbf{v}_M^T & 1 \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} n_x \\ n_y \\ n_z \\ b \end{bmatrix}}_{\beta} + \underbrace{\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_M \end{bmatrix}}_{\mathbf{e}}. \tag{2.4}$$

The unknown parameter β contains the normal vector and an intercept b capturing any offset at location $\mathbf{p}$ such as the one indirectly due to ambient light. The observation vector ζ consists of normalized intensity values at the collocated position $\mathbf{p}$ from images #1 to #M. The data matrix $\mathbf{X}$ is composed of vectors of incident directions, and the noise from measurement and/or modeling is modeled by a zero-mean error vector $\mathbf{e}$.

2.5.3 Proposed Method

Fig. 2.8 is a block diagram for the proposed authentication system. To authenticate a given test surface patch, $M > 4$ photos should be taken under flashlight. Each photo is processed to extract, warp, and register the captured patch to a grid of

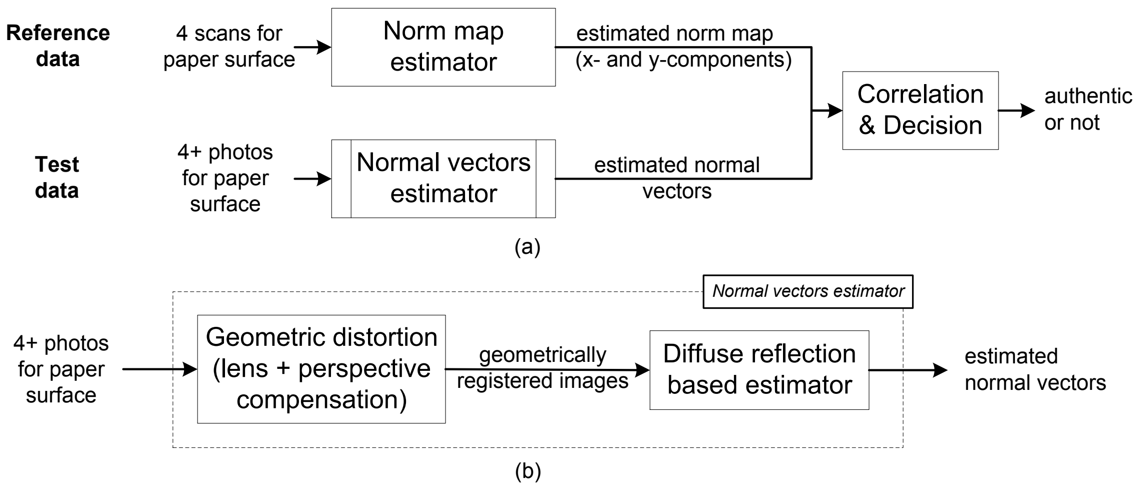

Reference data
4 scans for paper surface
Norm map estimator
estimated norm map (x- and y-components)
Correlation & Decision
authentic or not
Test data
4+ photos for paper surface
Normal vectors estimator
estimated normal vectors
(a)
Normal vectors estimator
4+ photos for paper surface
Geometric distortion (lens + perspective compensation)
geometrically registered images
Diffuse reflection based estimator
estimated normal vectors
(b)

Figure 2.8: (a) Block diagram for the proposed authentication system using mobile camera "captured" surface normal vectors. (b) Sub-diagram for the *normal vectors estimator* in (a).

200-by-200 pixels using the registration procedure outlined in Chapter 2.2.2. The resulting M registered patches with luminance nonuniformity are then processed by the diffuse reflection-based estimator proposed in Chapter 2.5.2. An estimated normal vector field is therefore obtained, and is used as an authentication feature for the surface patch. Its x- or y-component can be correlated with a reference to determine the authenticity using the hypothesis testing described in Chapter 2.2.3.

We treat the estimated norm maps from scanners as the reference, since they are reliable as discussed in Chapter 2.3.1 and relatively easy to obtain. More precise estimates of the norm maps can be obtained using microscopes. However, the benefit brought by the microscope with a much more controlled acquisition condition is marginal, and we will keep using norm maps from scanners as a reference.

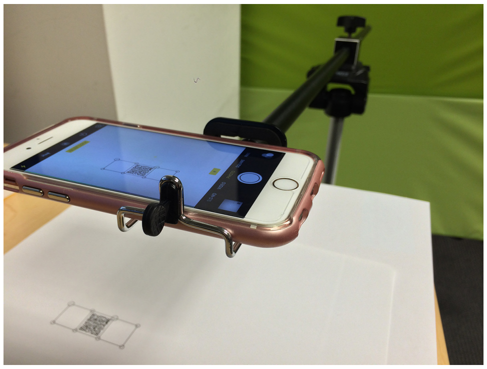


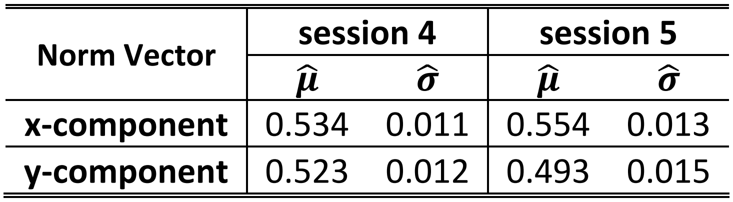
Figure 2.9: Setup for experiments conducted in Chapter 2.5.

2.5.4 Experimental Conditions and Results

Fig. 2.9 illustrates the experimental setup for capturing paper patches for estimating normal vector field using a mobile device. The mobile device was placed on a tripod and adjusted to be in parallel with the surface. The photos captured at the height of about $z = 11.4\,\text{cm}$ contain around 500 pixels along the edge/side for each patch in raw images, which are of enough resolution (5.25× more pixels than necessary) to generate the registered images of size 200-by-200. Detailed lighting conditions and models of mobile cameras are described individually for each capturing session.

One should note that the exact parallel configuration is not a required condition for our proposed method, and in this exploratory work, the parallel configuration was designed to avoid complications due to camera perspective. We later conducted additional experiments in which mobile cameras were held by hand and not exactly in parallel with the surface, and the results showed no degradation in authentication performance.

Table 2.4: Statistics for correlation values of matched cases (H_1) for images captured in a totally dark environment (Exp. 1).



Norm Vector	session 4		session 5	
	$\widehat{\mu}$	$\widehat{\sigma}$	$\widehat{\mu}$	$\widehat{\sigma}$
x-component	0.534	0.011	0.554	0.013
y-component	0.523	0.012	0.493	0.015

2.5.4.1 Totally Dark Environment

Two sessions, namely, Session 4 and Session 5 (*aka* **Exp. 1**), were independently captured using iPhone 6 at the same paper patch in a totally dark environment. Each session contains 20 camera captured images for the paper patch at 20 different locations indexed in Fig. 2.6(a).

For each norm map component and each session, we correlate the estimate from the mobile camera with the six estimates from two scanners (three slightly different norm maps for each scanner), and a set of six scores are obtained. A t-test is carried out over the group of scores to check if the correlation is significantly greater than 0.

The results in terms of the sample mean and sample standard deviation are shown in Table 2.4. It is revealed that either x- or y-component of Sessions 4 and 5 has a correlation around 0.5, and the t-tests show that all correlation values obtained are statistically significantly (p-value $< 10^{-9}$).

2.5.4.2 Environment with Ambient Light

We relax the totally dark assumption by investigating more realistic scenarios with the addition of the ambient light. Sessions 6–10 (*aka* **Exp. 2**) and Sessions 11–15 (*aka* **Exp. 3**) and were captured in a low-strength diffuse ambient light environment using iPhone 6 and iPhone 6s, respectively. Further, Sessions 21–25 (*aka* **Exp. 4**) were captured in an environment with ambient light at the strength of indoor offices using iPhone 6s.

In addition to the result obtained in Exp. 1 that the correlation achieved using norm maps from mobile camera is significantly greater than 0, we would like to further measure quantitatively the discrimination capability that can be achieved in terms of the ROC curve $\big(P_F(\tau), P_M(\tau)\big)$ and/or more compactly, the equal error rate (EER), as outlined in Chapter 2.2.3.

For the rest of this chapter, each session will generate only one correlation value in each normal vector component, and the value is calculated by averaging over the six scores that can be computed from correlating with the slightly different versions of the reference norm maps. This approach is an effort towards reducing the effect of the inaccurate norm map estimates used as references at the service provider side, without adding burden to users during the verification process.

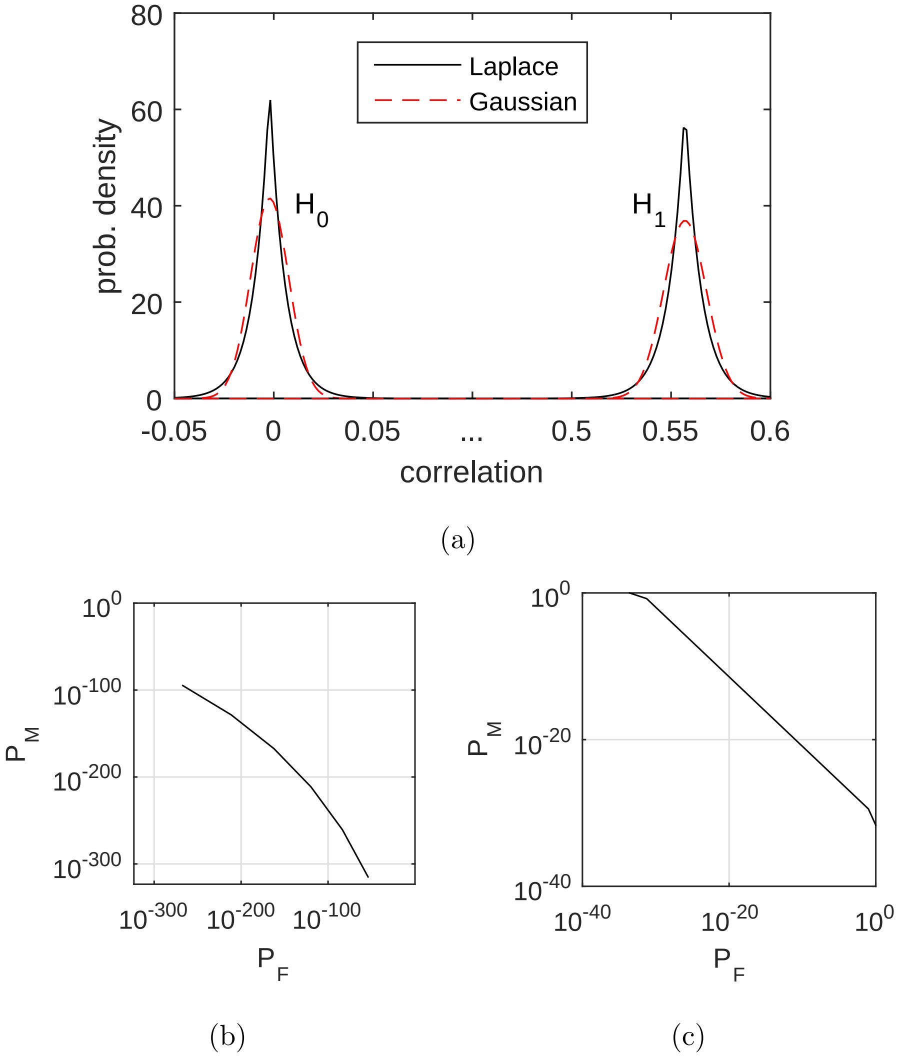
Fig. 2.10(a) shows the estimated PDFs for the matched (H_1) and unmatched (H_0) cases for Exp. 2. Under the acquisition condition for Exp. 2, the correlation values do not contain outliers and are distributed around a certain value. We therefore can consider they are sample points drawn from some probability distribution,

Table 2.5: Discrimination capability in EERs and corresponding statistics for images captured in environments with ambient light.

	Match (H_1)			No match (H_0)			EER	
	$\hat{\mu}_1$	$\hat{\sigma}_1$	$\hat{\lambda}_1$	$\hat{\mu}_0$	$\hat{\sigma}_0$	$\hat{\lambda}_0$	Gau.	Lap.
Exp. 2	0.557	0.011	122.8	−0.002	0.010	129.5	10^{-156}	10^{-16}
Exp. 3	0.532	0.015	97.4	−0.004	0.011	113.4	10^{-94}	10^{-13}
Exp. 4	0.528	0.012	106.4	−0.004	0.012	103.9	10^{-109}	10^{-13}

and we use this modeling assumption to help extrapolate the tails of PDFs and ROC curve. We select Gaussian and Laplace distributions to model the cases that the true distribution has light versus heavy tails, respectively. Detailed discussion on this modeling can be found in Chapter 2.7.5. Using the simple thresholding rule, we draw the ROC curves in Figs. 2.10(b) and (c). The discrimination capability measured in EER is 10^{-156} by assuming correlation is Gaussian distributed and 10^{-16} by assuming correlation is Laplacian distributed.

We list in Table 2.5 the discrimination capability for Exp. 2–Exp. 4 measured in EERs, and the detailed statistics that the EERs are calculated from, namely, the sample mean, $\hat{\mu}_i$, the sample standard deviation, $\hat{\sigma}_i$, the maximum-likelihood estimates for the rate parameter of the Laplacian distribution, $\hat{\lambda}_i$. As revealed by Table 2.5, the authentication performances are similar with different strength levels in ambient lighting and with different capturing devices (iPhones 6 and 6s). The high authentication accuracy and the flexible image acquisition procedure make the proposed method a promising technology to be deployed in a practical working environment. In addition to the above authentication performances that are measured



(a)

(b)

(c)

Figure 2.10: (a) Estimated PDFs of correlation for Exp. 2 (Sessions 6–10 & iPhone 6), (b) ROC curve by assuming correlation is Gaussian distributed, (c) ROC curve by assuming correlation is Laplace distributed.

for one acquisition condition per experiment, it is also beneficial in future work to measure the performance in a single experiment containing a variety of practical acquisition conditions.

2.5.4.3 Mobile Cameras of Other Brands

The investigation in this section has mainly used the cameras of the iPhone series for exploring the possibility of estimating the normal vector field of paper surface. We also carried out experiments using mobile cameras of other brands such as Samsung Galaxy Alpha. After obtaining the estimated normal vector field, we correlated the x- or y-component with the reference norm maps provided by scanners. The sample mean of correlation for matched cases (H_1) is around 0.23 with similar sample variance as in the experiments for iPhones. The smaller mean value compared to that of the iPhone cameras, 0.53, may be due to the fact that the flashlight of Samsung Galaxy Alpha is not so bright as those of iPhone cameras. The authentication performance measured in EER ranges from 10^{-22} to 10^{-6}. The EER results suggest satisfactory performances by the proposed method, and an effective decision strategy is to adjust the decision thresholds differently for Samsung Galaxy Alpha and for iPhone cameras considering their different PDFs of correlation under H_1.

2.5.5 Comparison with Prior Work

In Table 2.6, we summarize the performances of the proposed methods and prior art as discussed in the previous sections of the chapter. Our proposed image based method using mobile camera with flash has similar performance as the work in [15] that uses industrial camera and a semi-controlled light, as we created a semi-controlled light using flashlight and captured the test and reference images at the same position relative to the physical patch. The proposed image based method

Table 2.6: Authentication Performances of the Proposed Methods and Prior Art

Feature		**Modality**	**Lighting**	**Flexibility**	**Performance**
Type	**Detail**				**EER**
Image	pixel value	Industrial camera, Voloshynovskiy *et al.* [15]	semi-controlled	no	10^{-4}
	SIFT descriptor	Mobile camera, Diephuis *et al.* [17]	uncontrolled	yes	10^{-2}
	pixel value	**Mobile camera** (proposed in Chapter 2.4)	**semi-controlled**	**no**	$\mathbf{10^{-5}}$ **to** $\mathbf{10^{-3}}$
Norm map	seeded hash	Scanner, Clarkson *et al.* [14]	fully controlled	no	10^{-130} to 10^{-15}
	surface normal direction	**Mobile camera** (proposed Chapter 2.5)	**semi-controlled**	**yes**	$\mathbf{10^{-109}}$ **to** $\mathbf{10^{-13}}$

outperforms the method in [17] that uses robust point features, suggesting that a favorable lighting condition is a more important factor than a robust image processing technique.

Inspired by the success of various methods designed to use controlled lighting (such as the method proposed in Chapter 2.4 and the work in [15]) or inherently used controlled lighting [14], we have explored a semi-controlled lighting condition with the help of the flashlight of mobile cameras. The proposed norm map based method significantly outperforms all image appearance based methods. Although it performs slightly worse than the case using scanner as the acquisition device [14], the flexibility of the mobile device modality makes the proposed method more practical for ubiquitous deployment such as counterfeiting detection by end consumers.

2.6 Perturbation Analysis on Discriminability

This section analyzes the performance of the method proposed in the previous section under perturbations. We do not consider controllable factors that can potentially be taken care of at the service provider side, such as the number of norm maps used as references, and whether or not lens distortion should be compensated on the query images. Instead, we focus on the factors that are uncontrollable, such as the inaccuracy of the estimated camera locations, and the factors that may increase the burden to the users in the verification process, such as the number of flash images that users need to shoot in each session of verification.

2.6.1 Precision of Estimated Lens Location

The incident light direction $\mathbf{v}_i$ in Eq. (2.4) is an essential quantity for estimating the normal vector field. Its value is directly related to the camera location that may be imprecisely estimated. In this part, we first quantify the inaccuracy of the camera location estimate, and then perturb the camera location when calculating $\mathbf{v}_i$ to examine how the authentication performance will be affected.

Inaccurate Camera Location The standard deviation of location offset indicates how far away the estimated camera locations are from the true locations in a statistical sense. The true locations of the camera/lens were manually recorded while Sessions 4–10 were captured, each containing 20 images. Together with estimated locations calculated from the projection matrix (connecting the world and image coordinate systems), the 3-D location offset was obtained. For each image, the location offset is a vector containing quantities in x-, y-, and z-directions. The standard deviation for x-, y-, and z-directions were 1.86 mm, 2.16 mm, and 0.84 mm, respectively, when the camera was placed at the height of about $z = 11.4$ cm.

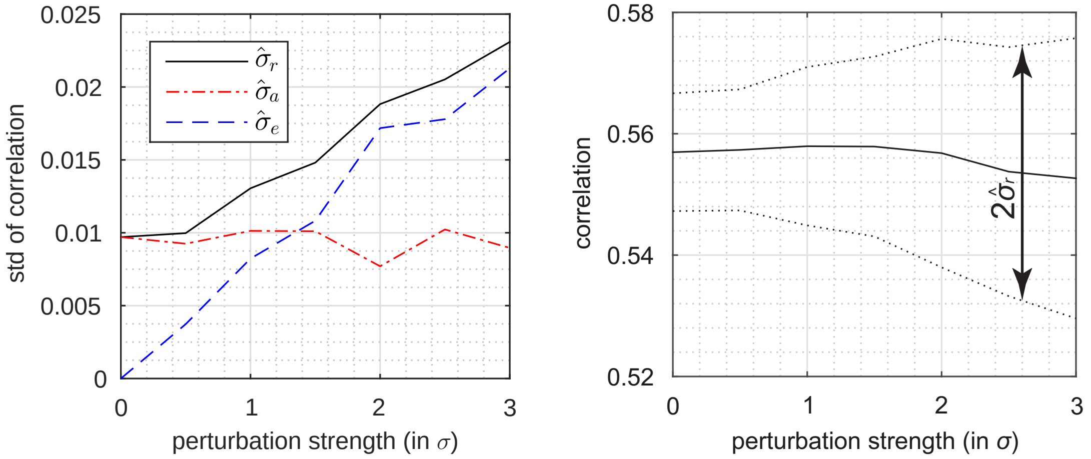
Performance Drop Under Perturbation With the knowledge of the amount of inaccuracy of camera location estimates, we can examine how authentication performance will be affected by adding a reasonable amount of perturbation. We chose $\sigma_x = 2\,\text{mm}$, $\sigma_y = 2\,\text{mm}$, and $\sigma_z = 0.9\,\text{mm}$ as the unit standard deviation in each direction, and scaled them by a list of scalars $[0, 0.5, 1, 1.5, 2, 2.5, 3]$, in which "1" corresponds to the nominal strength we obtained above. The larger the scalar is, the stronger the perturbation will be added.

For each perturbation level, a self-contained sub-experiment was carried out. The sub-experiment was carried out using the images from Sessions 6–10. For each session, the estimated camera location would be biased for 20 times by different location offset vectors that were independently drawn from the distribution of the current perturbation level. The resulting 5×20 correlation values are expected to have an increased variance due to the additional perturbation.

We analyzed the results of the sub-experiments using a *random effect model* [25] in order to reveal quantitatively the effect on the correlation value due to the additional perturbation and different sessions. The correlation r_{ij} obtained in the jth random trial of the ith session of the sub-experiment is assumed to be a summation of the mean correlation value μ (an unknown but fixed parameter), the zero-mean random effect a_i of the ith session, and the remaining error e_{ij} at the perturbation strength level of the current sub-experiment:

$$\begin{aligned} r_{ij} = \mu + a_i + e_{ij}, \qquad & i = 6, \cdots, 10, \\ & j = 1, \cdots, 20, \end{aligned} \tag{2.5}$$

where $a_i \sim N(0, \sigma_a^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$, and they are assumed to be jointly inde-



(a) (b)

Figure 2.11: (a) Estimated standard deviation of the correlation $\hat{\sigma}_r$ and its decomposition $\hat{\sigma}_a$ and $\hat{\sigma}_e$ for correct matches as a function of inaccuracy of lens location estimate, and (b) estimated mean correlation with two-sigma wide performance region, $(\hat{\mu} - \hat{\sigma}_r, \hat{\mu} + \hat{\sigma}_r)$.

pendent. Note that the variance of r_{ij} is composed of those of a_i and e_{ij}, namely, $\sigma_r^2 = \sigma_a^2 + \sigma_e^2$. We are interested in the values of the modeling parameters μ, σ_a, and σ_e, and the analytical expressions of the maximum-likelihood estimators are discussed in textbooks on the random effect model [25].

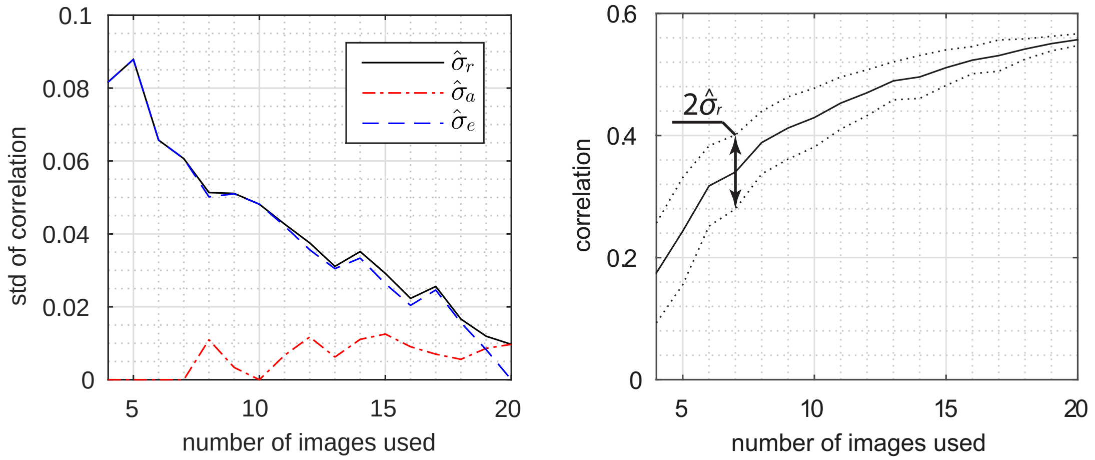
We obtained a distinct set of parameter estimates for each sub-experiment corresponding to a certain perturbation level, and plotted them with respect to the perturbation level accordingly. Fig. 2.11(a) shows a near-constant effect (quantified by $\hat{\sigma}_a$) for session, and an increased effect of perturbation (quantified by $\hat{\sigma}_e$) as the perturbation level increases. Fig. 2.11(b) shows the mean correlation against the perturbation level with two-sigma wide performance region. Both figures reveal that the resulting increase of perturbation is small compared to the mean correlation value, with $\hat{\sigma}_r < 0.014$ when the perturbation strength is at the nominal level, and $\hat{\sigma}_r < 0.024$ even when the perturbation strength is 3 times of the nominal level.

2.6.2 Number of Images for Normal Vector Field Estimation

We now consider the effect of the number of available images on the estimation of the normal vector. Recall in the main experiment, we used $M = 20$ images to estimate the normal vector field, which may not be very user friendly with a large number of light flashes in a short period of time. We varied the number of images from $M = 20$ down to $M = 4$, and carried out a self-contained sub-experiment similar to those in the last subsection using the images from Sessions 6–10. For each session, 20 subsets of the available images were selected and their correlation values are examined at the current perturbation level. The resulting 5×20 correlation values are expected to have an increased variance due to fewer images used.

Regarding the selection of the subsets of images, one should identify whether the available image set contains extremely "bad" ones towards the estimation of the normal vector and correlation, which should be carefully handled in the selection process. We tried to identify "bad" images using the following two criteria: i) the fitting error in the model of Eq. (2.4), and ii) the correlation improvement of when excluding an image. No "bad" image was identified out of the 20 available images, and we therefore constructed subsets of images by uniformly random selections from the indices $1, \cdots, 20$.

We analyzed the results of the sub-experiments using the same random effect



(a)

(b)

Figure 2.12: (a) Estimated standard deviation of the correlation $\hat{\sigma}_r$ and its decomposition $\hat{\sigma}_a$ and $\hat{\sigma}_e$ for correct matches as a function of number of images for norm map estimation, and (b) estimated mean correlation with two-sigma wide performance region, $(\hat{\mu} - \hat{\sigma}_r, \hat{\mu} + \hat{\sigma}_r)$.

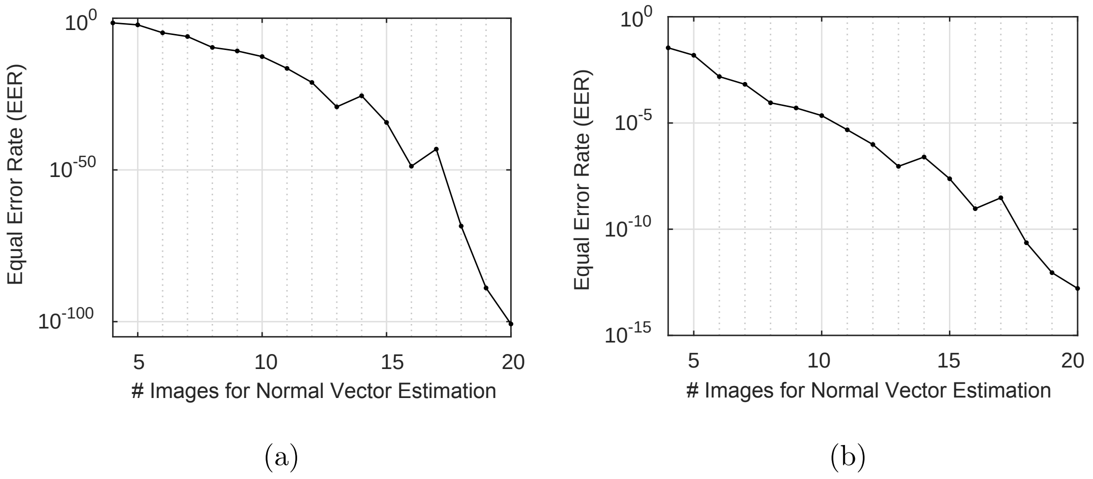
model as in the last subsection to reveal quantitatively the effect of having fewer images for the normal field estimation. We obtained a distinct set of parameter estimates for each sub-experiment corresponding to a certain number of images, and plotted them accordingly. Fig. 2.12 (a) shows, as expected, a constant effect for session, and an increased effect of perturbation as fewer images were used. The value of $\hat{\sigma}_r$ reaches almost 0.1 when the number of images used is reduced to 4. Fig. 2.12(b) shows that the mean correlation can drop to below 0.2 and the rate of the drop accelerates as the number of images reduces. Both figures reveal that the number of images used for norm map estimation can significantly affects the correlation.

2.6.3 Perturbation Factors Combined

The perturbation analyses in the above two subsections reveal that the number of images used for norm map estimation dominates the correlation value, compared to other factors such as the capturing session setup and the accuracy of the estimated camera location.

We now evaluate the discrimination capability in terms of EER by considering all possible factors investigated above. The EER will be plotted against the dominant factor, namely, the number of images. Such other remaining factors as the session setup, and the inaccuracy of the estimated camera location will be taken into consideration by boosting the overall variance in their respective amount estimated earlier in this chapter.

The two plots in Fig. 2.13 show the EER as decreasing functions of the number of images under Gaussian and Laplacian models, respectively. The results show that in order to obtain an EER of 10^{-4}, one should on average acquire at least 6 flash images if the correlation follows a light-tailed Gaussian distribution. In contrast, if the correlation follows a heavy-tailed Laplacian distribution, one should on average acquire at least 8 flash images. More discussions on modeling the PDFs of correlation using the Gaussian versus the Laplacian can be found in Chapter 2.7.5.



(a) (b)

Figure 2.13: Discrimination capability in terms of EER taken into consideration of all factors as a function of number of images for (a) Gaussian, and (b) Laplacian distributed correlation values.

2.7 Discussions

2.7.1 Interpretation of Norm Map From Low Resolution Images

When camera's capturing resolution is high enough, the area covered by each pixel is relatively flat, and the normal vector assigned to the pixel represents the physical surface direction of the area. The collection of the normal vectors therefore serves as a fingerprint for the paper surface.

When the resolution is lower than the aforementioned scenario, however, is the normal vector still a meaningful quantity? Let us relate a high resolution image and its low resolution version by a virtual 2-D low-pass filter with coefficients $\{w_i > 0 \,|\, \sum_{i=1}^{N} w_i = 1\}$, where i is a location index linearized from a 2-D index pair and N is the number of pixels covered by the filter. A pixel value u in the low resolution image is therefore the weighted sum of N pixels each with intensity $\mathbf{n}_i^T \mathbf{v}_i$ of the high-resolution image, where $\mathbf{v}_i$ and $\mathbf{n}_i$ are the directions of the incident light and normal vector at location with index i, respectively. Hence,

$$u = \sum_{i=1}^{N} w_i \cdot \mathbf{n}_i^T \mathbf{v}_i \approx \left(\sum_{i=1}^{N} w_i \mathbf{n}_i \right)^T \mathbf{v} = \bar{\mathbf{n}}^T \mathbf{v} \tag{2.6}$$

where $\mathbf{v}$ is the direction of the incident light for the pixel in the low resolution image. The approximation can be justified because in a small neighborhood, the direction of incident light is almost constant, *i.e.*, $\mathbf{v} \approx \mathbf{v}_i$. The term enclosed in the parentheses immediately on the RHS of the approximation sign can be regarded as the reflected intensity in a larger area with an averaged direction $\bar{\mathbf{n}} = \sum_{i=1}^{N} w_i \mathbf{n}_i$. That is, the norm maps estimated from low resolution images can be considered as a downsampled norm map using the virtual filter $\{w_i\}$ that relates the high and low resolution images.

2.7.2 Effect of Motion Blur

Slight panning motion during the capturing process often results in blurred images. The effect of the panning motion can be modeled by a linear-spatial invariant filter. In a special case that the motion blur is the same for all images captured, the normal vector field will be blurred by the same filter of the motion blur per the propagation property discussed above in Chapter 2.7.1. A blurred normal vector field may lead to a lower verification rate. It would be interesting to study how fast the authentication performance will drop as the strength of the motion blur increases. In a general case that the motion blur is not consistent for all images captured, the lowpass filtering effect does not directly propagate to the normal vector field. In this case, a study on how the motion blur will change the normal vector field and its ultimate impact can be carried out. If motion blur turns out to be a major factor for lowering the authentication performance, one can consider applying blind deconvolution in the first place for deblurring the image before using them for authentication purpose.

2.7.3 Understanding the Physics of Paper Surface Reflection

In this subsection, we use a confocal microscope to obtain the 3-D structure of a paper surface as a topographic map. This "ground-truth" map helps us examine the linkage between the reflected image appearance and the physical structure of the paper surface.

Normal Vectors From Confocal Microscope We use a Leica confocal microscope (under the reflection imaging mode using 488 nm laser light) to obtain a topographic map with a per-sample resolution of $3\,\mu$m, $3\,\mu$m, and $5.7\,\mu$m in x-, y- and z-directions, respectively. For the square surface patches of edge length $2/3$ inch digitized to 200 pixels (*aka* working pixels), the area covered by each pixel contains about 796 pixels in the topographic map (*aka* confocal pixels).

We estimate the normal direction for each working pixel described as follows. Fig. 2.14 is a sample collection of topographic blocks with each showing the area covered by one working pixel. Our examination over the whole set reveals that most blocks were not flat because the scale of fibers is smaller than the area of a working pixel. Using the result from Chapter 2.7.1, we calculate a surface direction

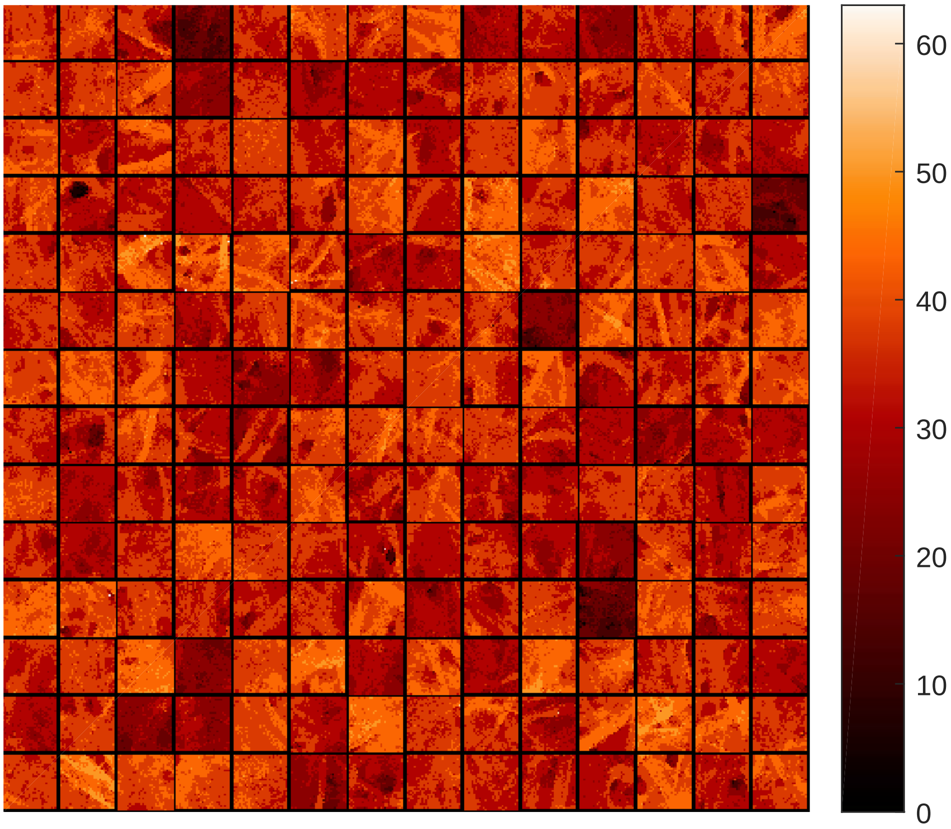


Figure 2.14: Sample collection of topographic blocks from copy paper of size 85 μm-by-85 μm by confocal microscope. The depth unit on color bar is also μm.

for each working pixel area by weighted averaging over the directions of all confocal pixels. Alternatively, we fit a plane to all confocal pixel locations and use the direction of the plane as an alternative estimate. These two estimates for the surface direction agree with each other with a correlation of 0.98, implying that the physical normal vectors are not sensitive to different definitions of direction and estimation algorithms, and are therefore reliable. Hence, the plane estimates are sufficiently good estimates for normal directions, and are considered as the physical ground truth in the experiments followed.

We examine the correlation of norm maps obtained from mobile camera and

Table 2.7: Correlation of Norm Maps with Ground Truth

Pair of Quantities	**Correlation**
Mobile camera *vs.* confocal (ground truth)	0.19
Scanner *vs.* confocal (ground truth)	0.28
[Reference]: Mobile camera *vs.* scanner	0.59

scanner with respect to the ground truth. The results are shown in Table 2.7. The non-zero correlation values imply that norm maps estimated by the scanner and mobile camera are indeed related to the ground truth, *i.e.*, the physical norm map obtained by the confocal microscope. However, the correlation values with the ground truth are low, around 0.2–0.3. This suggests that although the fully diffuse reflection model provides a surface norm estimation that is sufficiently discriminative for authentication purposes, the estimation may not be highly precise at the accuracy level of confocal microscopes.

Dominant Reflection Type In this part, we study the relative contributions of diffuse and specular components, with the help of the physical normal vectors from a confocal microscope. We first calculate a synthesized diffuse image $\mathrm{Im}_{\mathrm{d}}(\mathbf{p}) = \max\{0, \mathbf{n}(\mathbf{p})^T \mathbf{v}_i(\mathbf{p})\}$ and specular image $\mathrm{Im}_{\mathrm{s}}(\mathbf{p}) = \max\{0, \mathbf{v}_c^T \mathbf{v}_r\}$ using known quantities, without including the common effect of the light strength at location $\mathbf{p}$. Here, $\mathbf{n}$ is the surface normal vector, $\mathbf{v}_i$ is the incident light vector, $\mathbf{v}_c$ is the camera direction vector, and $\mathbf{v}_r$ is the specular reflection vector that can be represented as $\mathbf{v}_r = (2\mathbf{n}\mathbf{n}^T - \mathbf{I})\mathbf{v}_i$. We then regress 20 camera captured images $l_r(\mathbf{p})$ against the diffuse image and specular image, in order to obtain non-negative weights for diffuse and specular images scaled by the light strength, $\{w_d \cdot l^{(k)}(\mathbf{p}),\ k = 1, \cdots, 20,\ \forall \mathbf{p}\}$ and $\{w_s \cdot l^{(k)}(\mathbf{p}),\ k = 1, \cdots, 20,\ \forall \mathbf{p}\}$, respectively. Using non-negative matrix factorization with rank 1, we obtain estimates for w_d and w_s up to a multiplicative scalar. The ratio between the contributions of diffuse and specular components, w_d/w_s, is 5.82. This high ratio of nearly six-to-one reaffirms the diffuse reflection model in this chapter, and explains the excellent authentication performance in prior work [13, 14].

Generative Modeling Using the relative weights of diffuse and specular components in the reflection model of paper, we synthesize reflection images and examine their relationship with the camera captured images. We again use correlation as the similarity measure, and carefully remove the macroscopic trend of spatial intensity in order to avoid correlation inflation.

For 20 pairs of synthetic and camera captured images, we observe a statistically significant correlation of 0.13. This result from a generative modeling and the result of Table 2.7 from a discriminative modeling show that it is possible to connect the physical normal vectors to the surface appearance. Possible future directions to improve such connection include examining the role of diffraction, as well as the roles of transmitted and re-emitted light, as paper is not always fully opaque.

2.7.4 Robustness of the Norm Map

Practical deployment of the proposed scheme requires understanding on the robustness of the norm map under various conditions, and designing adaptive authentication algorithms when necessary. Clarkson *et al.* [14] conducted tests on scribbling with a pen and printing single-spaced text on around 10% of the test regions. They also tested the water treatment by using paper dried and ironed after being submerged. Their experiments demonstrated that the norm map is a robust feature under these conditions.

More investigations should be carried out on the resilience against tampering of the paper surface, such as scratching, folding, crumpling, and on the reliability of the physical structure over the time. Large scale tests for papers from different manufactures are beneficial to understand issues may arise in the deployment of the proposed technology. Below we discuss how to achieve resilience against the folding operation.

Resilience Against Folding Paper can be easily folded, resulting in a change of directions of those surfaces around the fold lines. In order to maintain a high correlation for true matches, the following strategies can be applied. The first strategy masks in correlation calculation those pixels whose surface directions are affected by folding. This method is intuitive but relies on the detection and segmentation of folded regions. As the distortion to the norm map field due to folding can be viewed as the addition of a slowly spatially varying trending surface, the second strategy is to apply detrending methods before calculating the correlation. For example, highpass filtering can be applied to remove the global trend. Such a highpass filter should be designed to properly reject the frequency components of the trending surface. Alternatively, parametric surfaces can be fitted to estimate the trending surface, and the resulting residue can be used to perform correlation. A practical challenge lies in the selection of a parametric surface that neither overfits nor underfits.

2.7.5 Considerations for Using Statistical Methods for Inference

In practice, the theoretical PDFs, $f_{\hat{\rho}|H_0}(\hat{\rho})$ and $f_{\hat{\rho}|H_1}(\hat{\rho})$, as well as the performance metrics, $P_D(\tau)$, $P_F(\tau)$, ROC, and EER derived from them, are not known *a priori*, and need to be estimated from the practical data obtained from experiments. One can construct normalized histograms or empirical probability mass functions (PMFs) for H_0 and H_1 as estimates for the true PDFs, and use the resulting PMFs to calculate the performance metrics. This approach using the real data can give estimates that are close to the true PDFs especially when the sample size is large. However, using PMFs as the estimates for PDFs leads to piecewise ROC curves and imprecise EER estimates, and the lack of distribution data in tails requires extremely large sample size to reveal the true performances around the two tail regions of the ROC curve. For example, for a 50-image dataset containing $\binom{50}{2} = 1,225$ possible pairs of images for verification, the smallest possible estimates for P_M and P_F are $1/50$ and $1/1225$, respectively, which may not precisely reflect the performance of system if the achievable rates are much smaller than 1/1225.

To alleviate the limitations of using the empirical PMFs for inferences, we can incorporate more modeling flavors by assuming the theoretical PDFs $f_{\hat{\rho}|H_0}(\hat{\rho})$ and $f_{\hat{\rho}|H_1}(\hat{\rho})$ follow some commonly seen distributions such as Gaussian and Laplacian. Adding this additional assumption has the advantage that the tails of PDFs and ROC can be better extrapolated, and EER can be calculated as a deterministic function of moments such as the mean and the variance. One should note that the accuracy of the extrapolated tails depends heavily on the assumption that the data would match with the assumed distribution. As sample points for tails are usually lacking for samples of a small size, it is reasonable to try a heavy-tailed distribution (such as Laplacian) to infer the lower performance bound, and to try a light-tailed distribution (such as Gaussian) to infer the upper performance bound. As can be recalled, we have calculated in Chapter 2.5.4 such performance bounds for both ROC curves and EERs.

2.8 Chapter Summary

In this chapter, we have investigated intrinsic microscopic features of the paper surface for authentication purposes. We have shown that it is possible to use the cameras and built-in flashlights of mobile devices to estimate the normal vector field of paper surfaces. Perturbation analysis shows that the proposed method is robust against inaccurate estimates of camera locations, and using 6 to 8 images can achieve a matching accuracy of 10^{-4} in EER under a lab-controlled ambient light environment. This finding can relax the restricted imaging setup in prior art, and enable paper authentication under a more casual, ubiquitous setting with a mobile imaging device. The proposed technique may facilitate duplicate detection of important and/or valuable documents such as IDs, and facilitate counterfeit mitigation of merchandise via detection of duplicated labels and packages.

3

High-Contrast, Extrinsic Features of Surfaces for Authentication

3.1 Chapter Introduction

In Chapter 2, we have studied using the norm map, a collection of directions of microscopic surfaces, as the intrinsic feature for authentication and counterfeit detection purposes. The choice of the norm map was to make the authentication more resilient to light conditions that can heavily affect the low-contrast visual features of the surface.

Another approach toward robust authentication is to consider using surfaces with high-contrast visual features. High-contrast features are optically distinct that light conditions such as flashlight and shadow do not change their visual representation significantly. Existing extrinsic features include: FiberTag [11] and Kinde Labels [12], which randomly distributes visible fibers on surfaces to make each patch unique; Ramdot [11], which uses randomly positioned color dots; and BubbleTag [11], which contains inherently randomly positioned bubbles in polymer.

The focus of this chapter is to understand for those high-contrast surfaces, how the authentication performance can be affected by inaccuracies within the verification algorithms, and by the design parameters of the features such as the size, shape, and density. To this end, we build physical high-contrast patches containing powder particles randomly distributed on paper surfaces and sealed under tapes. These lab-produced patches provide an initial and practical understanding of how a high-contrast feature based authentication system works, and reveal the priorities for subsequent explorations. We then use computer simulated surface patches to understand quantitatively how the design parameters and algorithm inaccuracies interact. We also propose a generative model of the high-contrast feature based authentication system to gain insights into the problem. The proposed generative model can predict the authentication performance of a candidate system with minimum simulation.

3.2 Background and Preliminaries

We use the same squared-shaped registration container and alignment algorithm discussed in Chapter 2.2 for the high-contrast patches.

The hypothesis testing framework discussed in Chapter 2.2 is used for authentication. Besides using the intensity of patch as a feature, we also use handcrafted features such as the spatial descriptors characterizing the locations of foreground objects in Chapter 3.4.3. Other features, such as random projections (RPs) [15], binarized RPs [15], and SIFT descriptors [16, 17], can be chosen to represent the surfaces. We leave the study of these features to future work.

Artifacts such as imprecise alignment, non-uniform lighting and glare due to light reflection are usually not accounted by the assumptions of theoretical models of the hypothesis. We remove these artifacts in the preprocessing stage to improve the coherence between the data and the modeling assumptions.

3.3 Comparative Study on Using Low-Contrast Feature

In this section, we show the limitation of using images of paper surfaces as the authentication feature. The visual appearance or the resulting digital image of a paper surface is an optical rendering of the 3-D microscopic structure created by inter-twisted fibers. The optical rendering follows the diffuse light reflection model in which the direction of the incident light plays an important role. An intensity gradient pattern in one image of the surface patch, *e.g.*, a darker-colored blob, may not be visible in another image of the same patch under a different lighting condition. We examine in this section how the images of surface patches comprised of these low-contrast patterns performance in authentication tasks.

Prior Work Voloshynovskiy *et al.* [15, 16] examined the imaging setup of using industrial cameras with a semi-controlled lighting condition—an overhead circular ring-shaped light source. The resulting images have consistent appearances due to the controlled lighting conditions, and as expected, achieved good authentication performance.

Mobile cameras were tested under uncontrolled ambient light [17], whereby the performance was considerably worse than the setup with semi-controlled light. One way to improve the authentication performance at the cost of increasing the design complexity of the authentication algorithm [17], is to identify for a test image all intensity gradient patterns whose strengths are above a certain threshold, and try to match as many patterns as possible with those in the reference image. As the acquisition conditions for the test and reference images are usually different, a high percentage of pattern matching between the two images is not guaranteed. This can lead to a lower authentication performance.

Our Experiments We take photos of our lab-produced paper surfaces [Chapter 2.2.2] using mobile cameras, and apply our precise alignment algorithm [Chapter 2.2.2] to extract pixels values within the square-shaped registration containers. The aligned patches captured using different acquisition devices are shown in Fig. 3.1 with non-uniform lighting and glare removed. Although a general similarity exists among the patches, small details differ a lot.

Experimental results show that for one patch surface, correlation values between images captured by different mobile cameras are not high: mean values are only around 0.2 to 0.4 for various combinations of acquisition devices for test and reference patches. We show in Fig. 3.2 the estimated PDFs, sample PMFs under H_0 and their bounds, and the resulting ROC bounds for the case that test and reference datasets are obtained using iPhone 6.

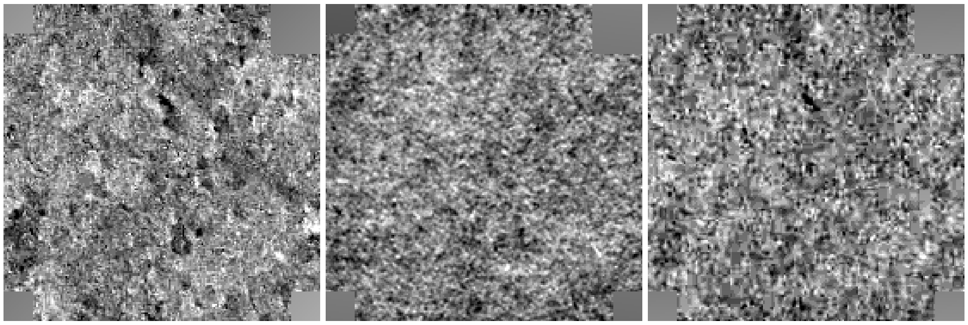


Figure 3.1: Images of a paper surface captured by scanner, iPhone 6, and Pantech Tablet, respectively. Contrast of the images are adjusted for display.

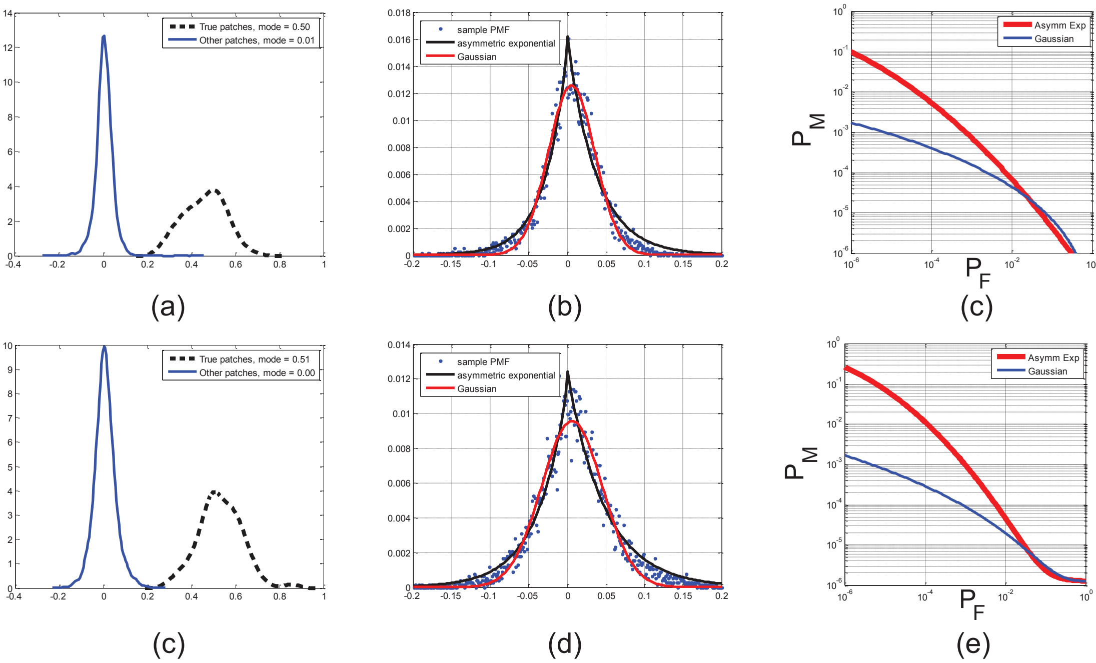


Figure 3.2: (a)(c) Estimated PDFs of $\hat{\rho}$, (b)(d) sample PMF under H_0 and its bounds (Gaussian and exponential distributions), and (c)(e) resulting ROC bounds. (a-c) for commonly found copy paper, (d-f) for light green card stock. At $P_f = 10^{-3}$, the miss rate P_m ranges from 10^{-4} to 10^{-3}. Both test and reference datasets are obtained using iPhone 6.

In Figs. 3.2 (b) and (d), we show the bounds for the data using two parametric distributions: Gaussian with light tails, and asymmetric exponential with heavy tails. The latter has the PDF of $f(x; \lambda_1, \lambda_2) = \lambda_1 e^{\lambda_1 x}\,\mathbb{I}[x < 0] + \lambda_2 e^{-\lambda_2 x}\,\mathbb{I}[x \geq 0]$. The well-fitted data in Figs. 3.2 (b) and (d) show that the sample PMFs are lower bounded by the Gaussian PDFs and are uppper bounded by the exponential PDFs. Using these two parametric distributions, we draw the performance regions in the ROC space as shown in Figs. 3.2 (c) and (e).

We estimate the PDFs with the following procedures. For the null hypothesis (unmatched cases), we fit both Gaussian and asymmetric exponential. For Gaussian, we assume zero mean and estimate the variance. For the asymmetric exponential distribution, we conduct parameter estimation for the positive part and negative part separately. For each part, we use an altered version of the method of moments to estimate the rate parameter λ. We note the exponential distribution has the following properties: $\mu = \lambda^{-1}$, $\sigma^2 = \lambda^{-2}$, and median $= \lambda^{-1}\ln(2)$. We equate these theoretical quantities to their corresponding sample statistics, and obtain three candidate estimates for λ. We then take the smallest one of the three as the final estimate, as it corresponds to the PDF that cover the largest amount of points in Figs. 3.2 (b) and (d). The resulting tails of the estimated exponential PDFs upbound sample PMFs well. For the alternative hypothesis (matched cases), we simply fit a Gaussian as the number of data points is far from adequate (fewer than the data points for the null hypothesis) to sensibly distinguish two distribution types.

The performance regions in the ROC plots of Figs. 3.2 (c) and (e) show that at 10^{-3} false-alarm rate, the miss detection rate ranges from 10^{-4} to 10^{-3}. This is barely acceptable in practical large-scale counterfeit detection applications. We see that even using newer mobile cameras such as iPhone 6 with improved acquisition quality, the authentication performance using the low-contrast visual appearance as a feature is limited.

3.4 Authentication Using High-Contrast Extrinsic Features

In this Chapter, we consider using surfaces with high-contrast visual features that are optically distinct that lighting conditions such as flashlight and shadow do not change their visual representation significantly.

3.4.1 Lab Produced Low-Cost Sealed Powder Patch

We create an experimental paper surface with dark flocking powder from craft stores as the foreground, to understand the performances of such surface with the randomly distributed visible fiber. The powder is randomly dropped on the paper surface to form a unique pattern, and the pattern is sealed by transparent adhesive tape. Each piece of flocking powder is short-bar–shaped. The design parameters of this powder surface include the density of the powder, and the spatial distribution. This lab produced surface [examples in Fig. 3.3] helps us understand the authentication performance of the surfaces as a function of design parameters.

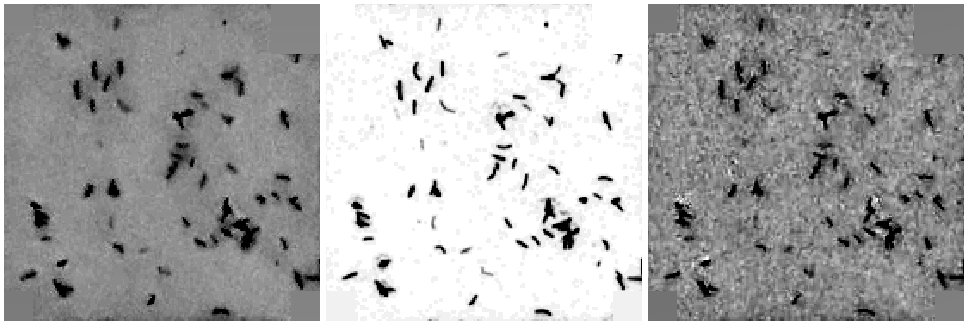


Figure 3.3: Images of a lab-produced powder surface captured by scanner, iPhone 6, and a Tablet, respectively. Contrast of the images were adjusted for display.

3.4.2 Performance Achieved Directly Using Images

Fig. 3.4 shows that once high resolution registration is achieved, the authentication performances for sealed powder surfaces are very good. We observe from experiments that better camera and higher powder density lead to larger margin between the densities of H_0 and H_1.

3.4.3 Performance Achieved Using Spatial Descriptors

In this part, we extract four spatial features, design distance measures, and construct statistics. We found that using merely a single feature—the locations of bars—can give good discrimination capability.

Feature Extraction The feature extraction algorithm is consist of i) combinations of morphological operators, ii) adaptive thresholding, iii) connected region search (breadth/depth first search, linear time of total number of image pixels), and iv) denoising using bilateral filter. The bilateral filter is used because it is capable of

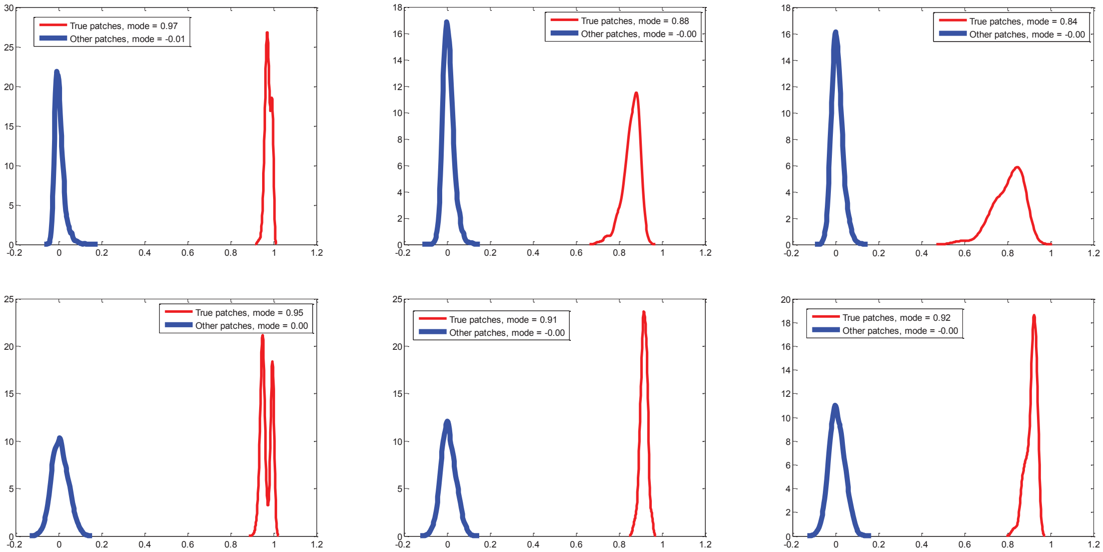


Figure 3.4: PDFs of $\hat{\rho}$ under H_0 (blue, thick lines) and H_1 (red, thin lines). 1st row: sparse powder, 2nd row: dense powder. Columns from left to right: scanner (test) vs. scanner (ref), iPhone 6 (test) vs. scanner (ref), and Pantech Tablet (test) vs. scanner (ref).

removing noise while preserving the foreground objects, especially those foreground objects that are slightly above the noise level.

The following information of each visual blob (most likely to be a short bar) is obtained for constructing a descriptor with four features: i) blob location, ii) blob area in pixel count, iii) length of major axis of a fitted ellipse, and iv) orientation of the fitted ellipse.

Fig. 3.5 shows annotated descriptors overlaid with the acquired images. Each detected blob is annotated with a dot showing the estimated center, and a line segment showing the major axis of a fitted ellipse. We observe that for high resolution acquisition devices like the scanner and iPhone 6, most blobs are correctly detected.

Pantech Tablet performs a little worse, but iPhone 3Gs give very bad results due to large background noise.

Distance Measure for Descriptor We design a distance measure that uses the blob location. We leave other information such as blob area, length of major axis, and orientation to future work.

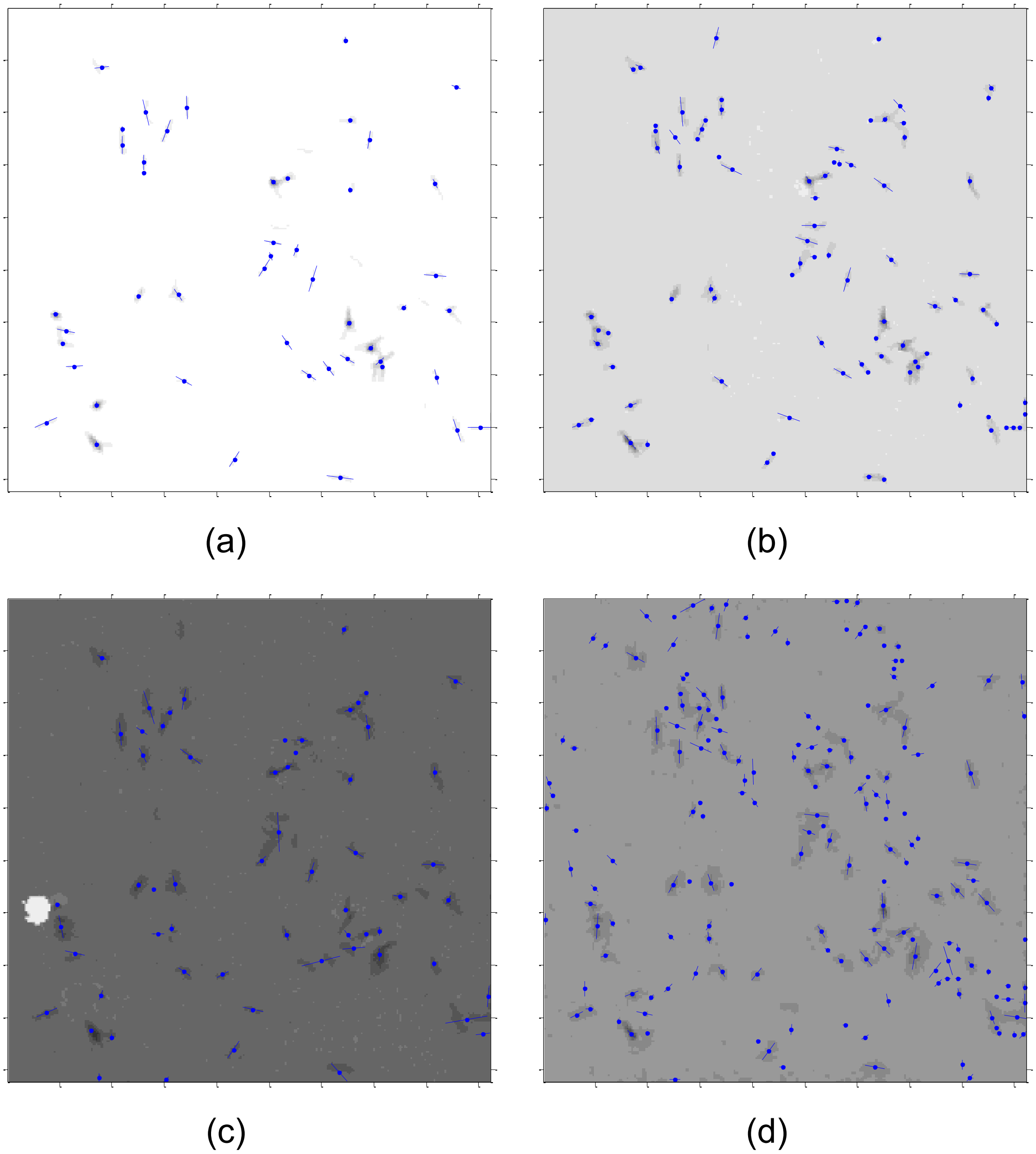
Suppose there are N detected blobs in the reference patch. For each detected blob in the reference patch, we examine at the test patch if there are blobs falling within the range of a ball of radius R pixels. In our experiment, we set R to 3. We set up three counters as follows:

- if no blob in the ball, then no match flag `flagMiss++`;
- if one blob, then correct flag `flagMatch++`;
- if more than one blob, then over match flag `flagOverMatch++`.

The sum of these three counters equals N. We define three reference patch based statistics:

1. `flagMissRatio = flagMiss/N`,
2. `flagMatchRatio = flagMatch/N`,
3. `flagOverMatchRatio = flagOverMatch/N`.

We also define a test patch based statistics, `unmatchedIndicesRatio`, the percentage that the detected blobs in the test patch that do not find matches in the reference patch.



(a)

(b)

(c)

(d)

Figure 3.5: Annotated Img#1 of db1, page 1 with descriptors. Each detected blob is annotated with a dot showing the estimated center, and a line segment showing the major axis of a fitted ellipse. Acquisition device: (a) scanner, (b) iPhone 6, (c) iPhone 3Gs, and (d) Pantech Tablet.

Experimental Results We apply every test statistic for authentication using images captured from four different mobile cameras. We found that among all proposed statistics, three have great discrimination capabilities, namely, `flagMissRatio`, `flagMatchRatio`, and `unmatchedIndicesRatio`. The PDF plots for the three statistics and four acquisition devices are shown in Fig. 3.6. The clear separation of the PDFs under two hypothesis reveals that the proposed statistics based on the blob location feature is powerful.

3.5 Perturbation Analysis

3.5.1 Analysis Using Real Images

We now investigate how the inaccurate alignment affects the discrimination capability when the image intensity is used as the authentication feature.

We introduce spatial perturbation before correlation of two aligned images is calculated. More specifically, a random perturbation vector $\Delta\mathbf{p}$ is added to the maker positions of aligned test images, whereas the marker positions of reference images are not changed. The distributions of the perturbation, *e.g.*, constant-value (*aka* fixed), constant-variance (*aka* random), and uniform (*aka* random), do not significantly affect the performance. We therefore report the results due to the constant-value perturbation. The corresponding $\Delta\mathbf{p}'$ added to the marker positions of unaligned test images are obtained using image registration algorithm described in Chapter 2.2.2.

The direction of $\Delta\mathbf{p}$ does not have an affect on the discrimination capabil-

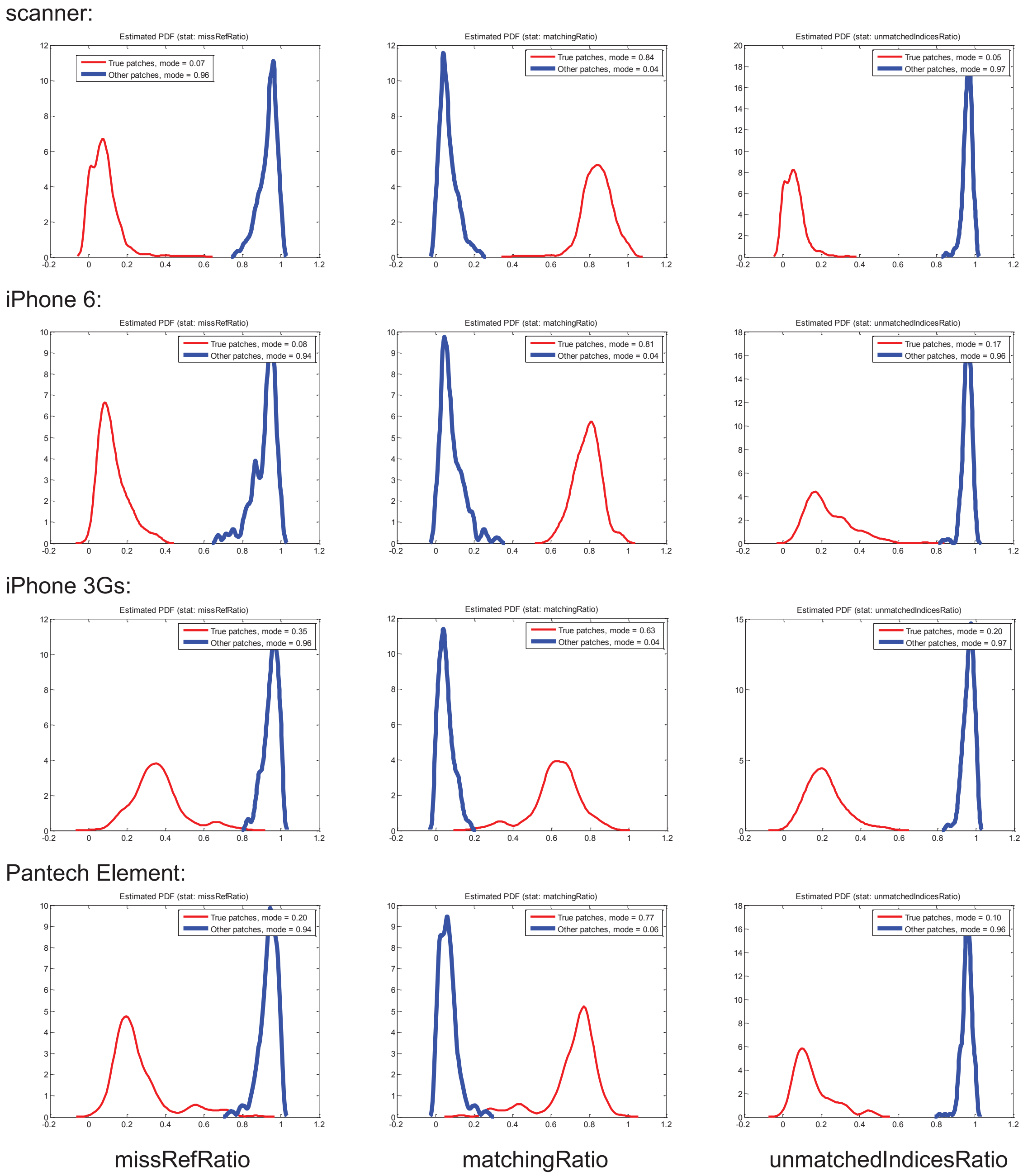
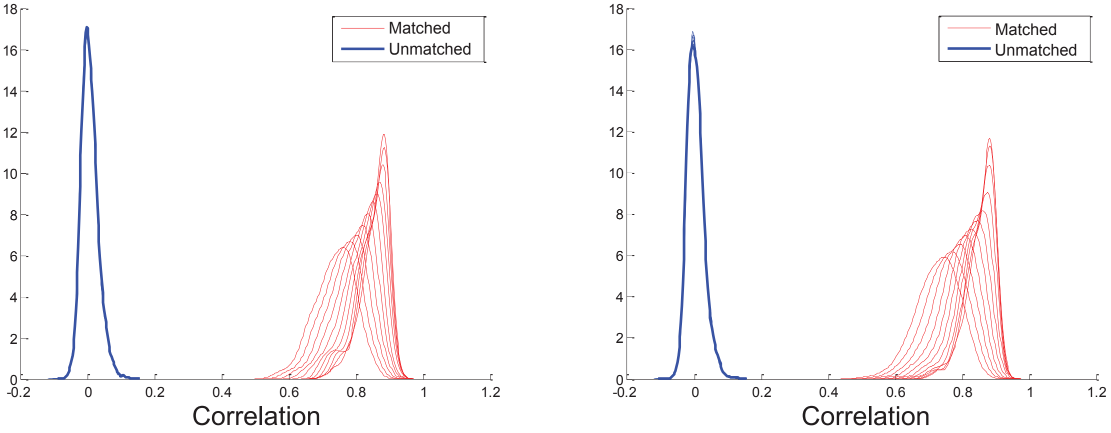


Figure 3.6: PDFs of $\hat{\rho}$ for matched (red, thin lines) and unmatched cases (blue, thick lines) for the three spatial features. Results are presented for four different acquisition devices.



(a) (b)

Figure 3.7: PDFs of $\hat{\rho}$ under H_0 (blue, thick lines) and H_1 (red, thin lines) for added alignment error ranges from 0 to 1 with stepsize 0.1. Error vector directions are up in (a) and right in (b). PDF under H_1 (red) moves to the left as the strength of perturbation increases. Data: iPhone 6 (dataset#1–3) vs. scanner (dataset#1), page 1.

ity, according to our experimental results on six directions (horizontal, vertical, up, down, diagonal, and anti-diagonal). This is expected because orientations of foreground objects are uniformly distributed. We show two results in Figs. 3.7 (a) and 3.7 (b) when the direction of the perturbations are up and right, respectively. In the following experiments, we assume that the perturbation vectors are pointing to the right.

We show the error bar plot in Fig. 3.8 to visualize how the statistical performance changes as the alignment error increases. The horizontal axis is the level of perturbation in the unit of pixel we added before calculating the correlation. The red curves show the correlation for the matched cases (H_1), whereas the blue curves

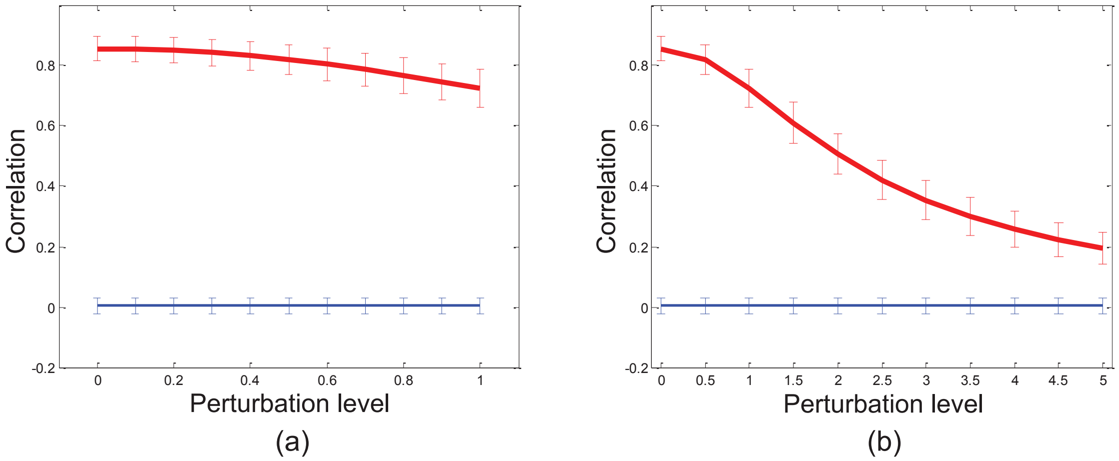


Figure 3.8: Error plots for average values of $\hat{\rho}$ under H_0 (blue, thin lines) and H_1 (red, thick lines) as a function of the strength of alignment error. (a) with a range of 0 to 1 pixels is a zoomed-in version of (b) with a range of 0 to 5 pixels.

show the correlation for unmatched cases (H_0).

It is observed that the distributions of unmatched cases is not affected by the perturbation, whereas the distributions of matched cases move to the left as the strength of the perturbation increases. This implies that the discrimination capability drops as perturbation increases. According to our data, the perfect separation of distributions ends as constant perturbation strength $\|\Delta\mathbf{p}\|$ increases to 3.2 pixels.

3.5.2 Analysis Using Synthetic Images

The 2-D appearance of BubbleTag [11] can be modeled as open circles. Gray scale version of Ramdots [11] can be modeled as filled circles/discs. We simulate BubbleTag and Ramdot surfaces by a set of foreground circles with different boundary widths. Foreground circles with random positions are drawn 600 ppi digitally, and

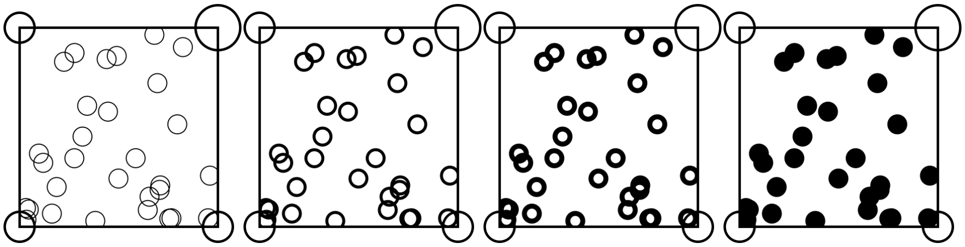


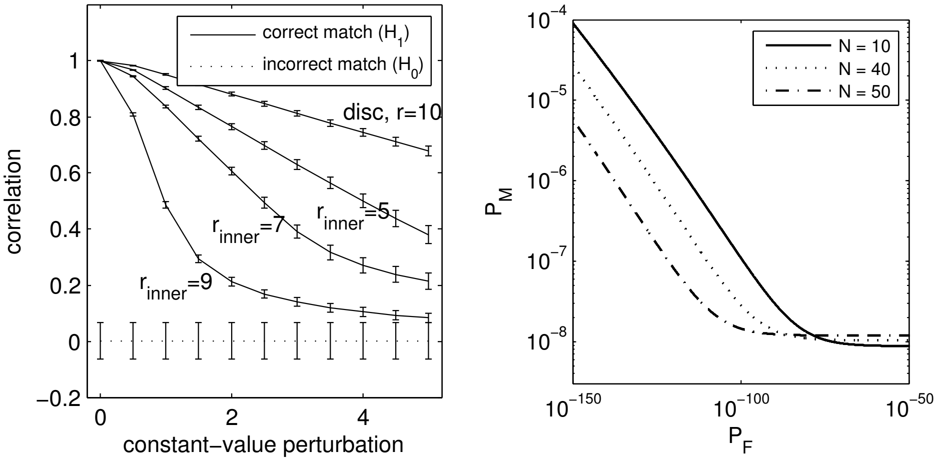
Figure 3.9: Simulated BubbleTags in 2D view with design parameters $r_{\text{out}} = 10$ and $r_{\text{in}} = 9, 7, 5, 0$.

then "recaptured" via antialiasing and downsampling to 300 ppi. Printing and scanning noise is assumed to be negligible. Misalignment is added in the form of constant-value perturbation.

Fig. 3.9 shows circles with same outer radius 10 and different inner radix 0, 5, 7, and 9; each patch has 30 circles. Our perturbation result shown in Fig. 3.10 (a) reveals that filled circles as foreground objects are less sensitive to alignment imprecision than unfilled ones, and circles with thicker boundaries are less sensitive than thinner ones. For a patch with circles of boundary width d, the convexity of the curve for averaged correlation under H_1 w.r.t. perturbation starts to change around perturbation level d.

Fig. 3.10 (a) shows that it is possible to have poor authentication performance when alignment is imprecise. Specifically, at perturbation strength equals 1, patches with circles of the thinnest boundary, *i.e.*, $r_{\text{inner}} = 9$, can have less than 0.5 correlation on average.

As alignment algorithms are usually capable of achieving subpixel precision, we consider the scenario that the length of displacement vector is uniformly dis-



(a) (b)

Figure 3.10: (a) Error bar plot of PDFs of $\hat{\rho}$ under H_1 for $r_{\text{out}} = 10$ and different r_{in}'s. The PDF under H_0 with the maximum std range is displayed as the worst case. (b) ROC curves for different number of foreground circles, N.

tributed between 0 and 0.5, and carry out experiments to understand the factors to authentication performance. We build on the case of circle inner radius = 7 [second patch in Fig. 3.9]: we fix the circle size, and change the density by changing number of circles for the surface. The result in Fig. 3.10 (b) shows a better performance for higher density.

3.6 Generative Model of High-Contrast Images

The experiments and simulations in the previous sections show that authentication performance depends on several factors, including alignment precision, and shape and density of appearance features. In order to go from qualitative understandings on factors affecting the authentication performance of surfaces to quantitative knowledge, it is beneficial to model the problem into several subproblems which can be solved analytically or numerically. This divide-and-conquer strategy give a better understanding in the the structure of the problem and the computation is minimal comparing to experimental and simulation approaches. We now briefly summarize how we approach the problem as follows.

We focus on a class of surfaces with isotropic foreground appearance features. To assess the authentication performance, the distributions of the test statistic—sample correlation coefficient $\hat{\rho}$—should be known under the null and alternative hypotheses. It is relatively easier to first approach the problem analytically by assuming the images of surfaces are binary. We also assume that the level of acquisition noise is not strong enough to flip any of the binary pixels. We further assume that the perspective matching error only happens in form of a global displacement. Hence, the description of the hypotheses are refined as:

H_0 (unmatched case): The acquired patch is a random patch, with the same design parameter Γ, *e.g.*, number of foreground objects and they are of the same shape, as the those in the reference patch.

H_1 (matched case): The acquired patch is identical to the reference patch, with global matching imprecision quantified by a random displacement vector.

The randomness of $\hat{\rho}$ under H_0 comes from the randomly distributed foreground objects, whereas the randomness of ρ under H_1 comes from the random matching imprecision.

The authentication performance under different choices of parameter values can be revealed using ROC curves for a general understanding of the performance, or using detection rate P_D for a specific α-level design constraint. A design parameter that maximized the performance can then be selected.

The correlation in both hypotheses are related deterministically to the number of black-to-white pixel flips, denoted as $\delta^{0\to1}$ or δ_m, and the number of white-to-black pixel flips, denoted as $\delta^{1\to0}$ or δ_n. Under the matched case H_1, we have $\delta = \delta^{0\to1} = \delta^{1\to0}$ due to isotropicity, where δ is related to the displacement vector quantifying the alignment imprecision. Under the unmatched case H_0, the joint distribution of $(\delta^{0\to1}, \delta^{1\to0})$ conditioned on one of a series of combinatoric cases can be analyzed and results can be obtained numerically. Simulated data are used to confirm the effectiveness of the model.

3.6.1 Decision Statistics as a function of (δ_m, δ_n)

When a binary test image is correlated to a binary reference image, the resulting sample correlation coefficient is a function of the number of foreground and background pixels M and N, and number of $0 \to 1$ flips, δ_m, and number of $1 \to 0$ flips, δ_n, needed for converting the reference image into the test image. Fig. 3.11 illustrates an example of pixel correspondences between two binary images and annotates the four variables defined above. It is not difficult to derive that the correlation as

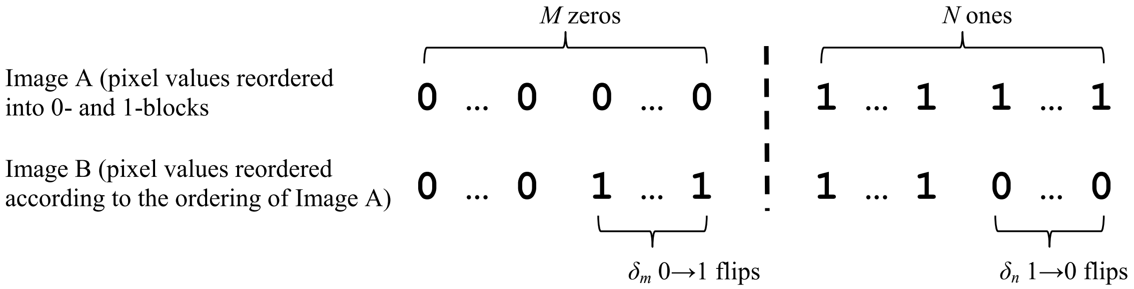

M zeros
N ones
Image A (pixel values reordered into 0- and 1-blocks
0 … 0 0 … 0
1 … 1 1 … 1
Image B (pixel values reordered according to the ordering of Image A)
0 … 0 1 … 1
1 … 1 0 … 0
δm 0→1 flips
δn 1→0 flips

Figure 3.11: Example of pixel correspondences between two binary images with annotated counts

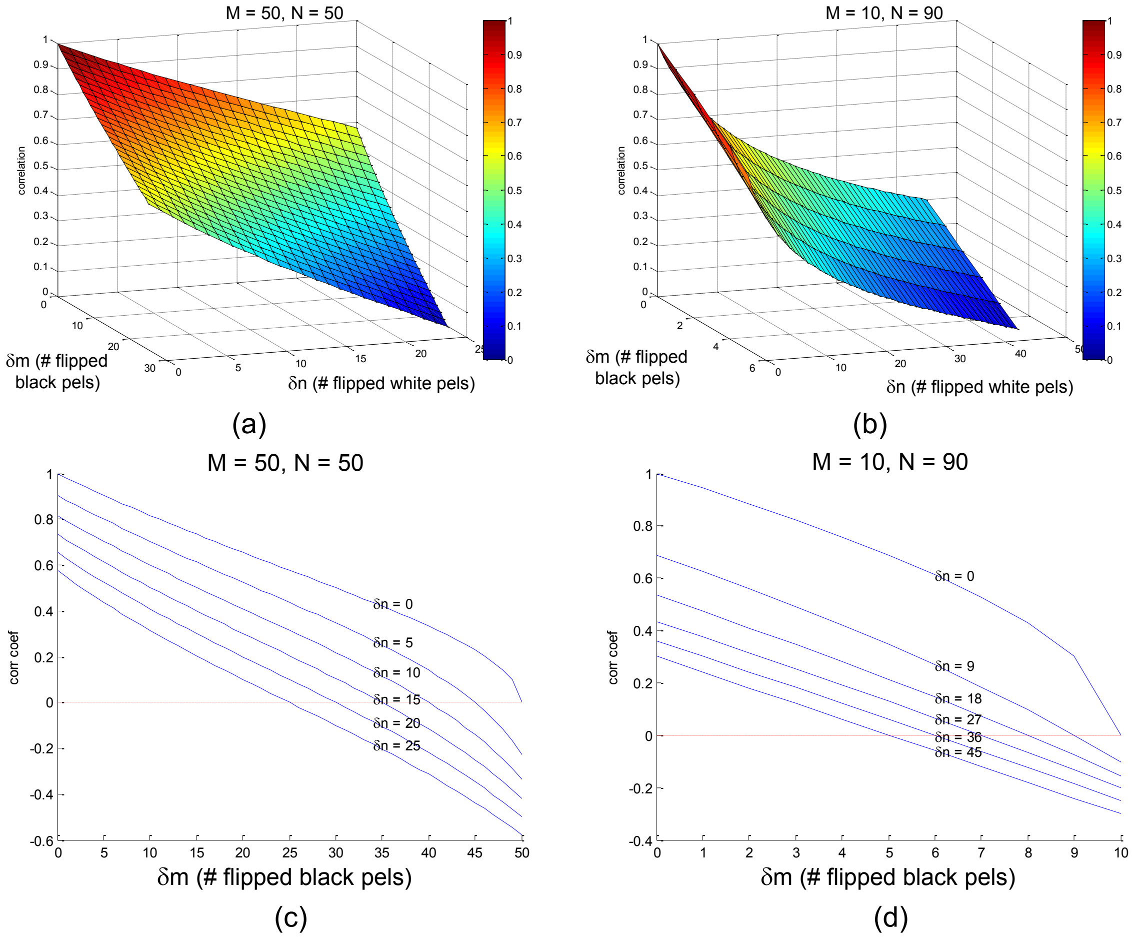
follows:

$$\begin{gathered}\hat{\rho}(\delta_m, \delta_n; M, N) = \frac{MN - (M\delta_n + N\delta_m)}{\sqrt{MN(M - \delta_m + \delta_n)(N - \delta_n + \delta_m)}} \\ 0 \leq \delta_m \leq M,\ 0 \leq \delta_n \leq N\end{gathered} \tag{3.1}$$

where $\hat{\rho} = 1$ if and only if $\delta_m = \delta_n = 0$, *i.e.*, two images are identical.

The symbols 0 and 1 representing the two binary pixel levels are not special of each other. For easy presentation, we name symbol 0's as foreground black pixels, and symbol 1's as background white pixels, to mimic scenarios that black dots lay on white paper.

Figs. 3.12 (a)-(b) reveal that, when the number of foreground pixels is the same as the number of background pixels (*aka* the balanced case), *i.e.*, $M = N$, the sample correlation drops almost linearly w.r.t. number of flips in black/white pixels. However, when the number of foreground pixels is less than the number of background pixels (*aka* the unbalanced case), *i.e.*, $M < N$, the sample correlation drops faster when white pixel flips. The same data (used above to draw the surfaces) are visualized in 1D parametric form, as shown in Figs. 3.12 (c)-(d). Again for the



(a) (b) (c) (d)

Figure 3.12: Sample correlation $\hat{\rho}$ as a function of δ_m and δ_n, with foreground-to-background ratio M/N (a)(c) $1:1$, and (b)(d) $1:9$

balanced case, we see the drop of correlation is almost linear w.r.t. number of flips of black and white pixels. For the unbalanced case, the drop of correlation is almost linear w.r.t. number of flips of black pixels (δ_m), but highly the drop is nonlinear w.r.t. number of flips of white pixels (δ_n, across the series of curves).

When two identical binary images are aligned off by a certain displacement vector (assume no foreground pixels are at boundaries), the number of flipped foreground pixels and the number of flipped background pixels are the same, *i.e.*,

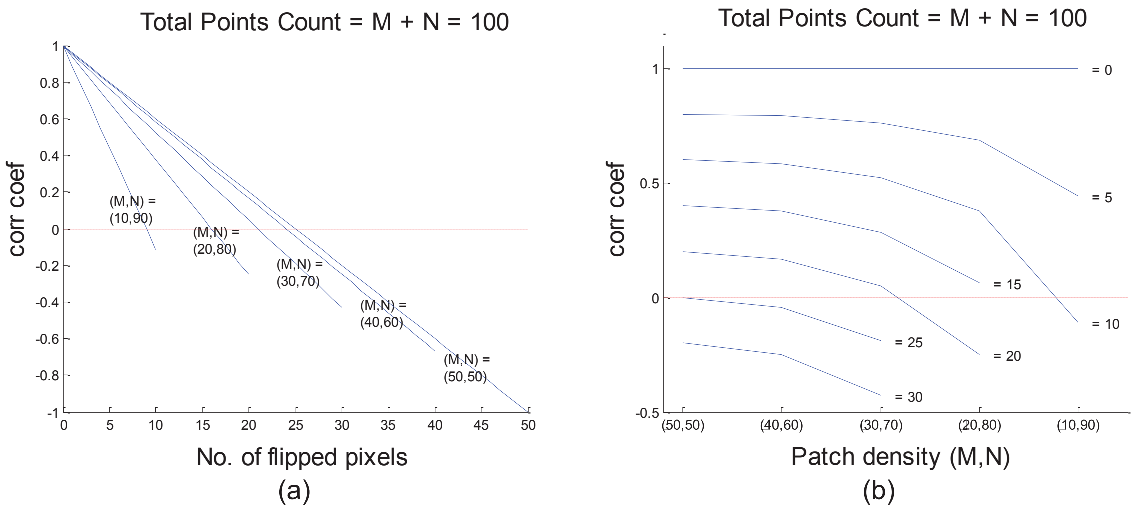


Figure 3.13: $\hat{\rho}$ of two identical images, *i.e.*, the autocorrelation, vs. the number of flipped pixels Δ.

$\delta_m = \delta_n$. Define the number of flips as Δ, the correlation of Eq. (3.1) is actually the autocorrelation of the binary images, and Eq. (3.1) is simplified to a linear function of Δ:

$$\hat{\rho}(\Delta) = 1 - \frac{M+N}{MN}\Delta, \qquad 0 \leq \Delta \leq \min(M, N). \tag{3.2}$$

Hence, we can condense the above figures into autocorrelation versus number of flips Δ parameterized by density [(M, N) pair represents the density: the small M is, the less density the binary image is], as shown in Fig. 3.13 (a). It is revealed that at a fixed alignment precision (*e.g.*, for now consider fixed Δ), the autocorrelation is better if the binary image is more balanced.

One note is that for the same level of mis-alignment, the number of flipped pixels can be different at the same density level: if all foreground pixels are singular, the number of flip reaches the maximum; if all foreground pixel form one

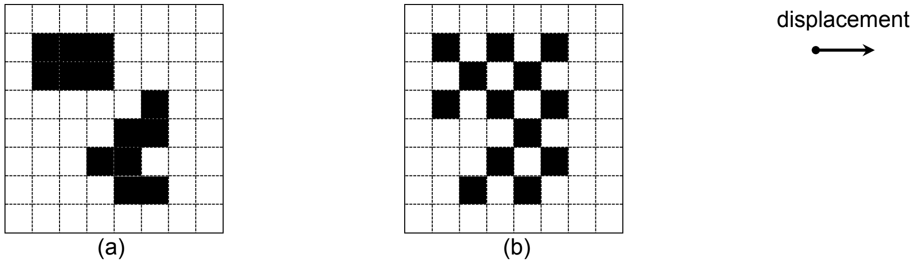


Figure 3.14: 8-by-8 binary images with 13 foreground pixels. The displacement vector (1,0) is pointing to the right. (a) $\delta_m = \delta_n = 6$, and (b) $\delta_m = \delta_n = 13$.

cluster, the flips only happen at the boundary pixels perpendicular to the direction of displacement vector so number of flips should be relatively low. The contrast of Figs. 3.14 (a) and (b) is such an example. This observation suggests that, less clusters can result in less flipped pixels.

To summarize, in order to obtain high autocorrelation for correct matches when only mis-alignment exists, one should design the patch so that the amount of foreground and background pixels should be similar, and foreground objects should form as less clusters as possible.

Note that the above is strategy is to enhance the correlation for correctly matched binary images/patches. However, the correlation of incorrectly matched patches will increase in very extreme cases, for example, there is only one foreground cluster. This scenario can be avoided in *ad hoc* by restricting the number of foreground clusters to be greater than a certain value, say, 10. The single-cluster patch is a highly atypical patch, and the number of atypical patches is negligibly small among all possible patches, and the restriction imposed should be able to generate typical ones that result in zero correlation for the incorrectly matched cases.

3.6.2 Derivation for Joint Distribution $f_{\Delta_m,\Delta_n|H_0}(\delta_m, \delta_n)$ Using Isotropic Foreground Objects

3.6.2.1 Matched Case, H_1

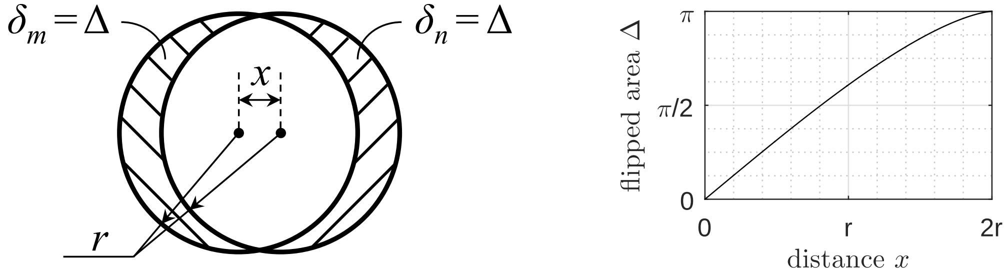
Equation (3.2) linearly relates the correlation $\hat{\rho}$ with the number of flips Δ (or flipped area for an image in a continuous spatial domain), and the remaining task is to relate flipped area with the alignment imprecision.

When the alignment imprecision is restricted to merely the global displacement $\mathbf{d}$ and the foreground object is isotropic, the flipped area is a deterministic function of the distance between the two objects, $x = \|\mathbf{d}\|$. For a disc of radius r as shown in Fig. 3.15 (a), by applying geometry it is not difficult to obtain the flipped area as a function of the distance of the two centers, x:

$$\Delta(x) = \left\{1 - \left[\frac{2}{\pi}\cos^{-1}\left(\frac{x}{2r}\right) - \frac{x}{\pi r}\sqrt{1-\left(\frac{x}{2r}\right)^2}\right]\right\}\pi r^2, \qquad 0 \le x \le 2r. \tag{3.3}$$

The relationship is shown in Fig. 3.15 (b). It is observed that when the distance x is relatively small compared to the radius r, the relationship is approximately linear.

When a particular alignment algorithm is used, the alignment imprecision can be profiled using synthetic data into an empirical probability distribution for the length of displacement vector x. For the $^1/_3$-pixel precision alignment algorithm we used in the experiments, the alignment imprecision can be assumed to be uniformly distributed between 0 and $^1/_3$.



(a) (b)

Figure 3.15: (a) Two identical discs with radius r separated by distance x has two pieces of non-overlapping area δ_m and δ_n, where $\delta_m = \delta_n = \Delta$. (b) The deterministic relationship between Δ and x.

3.6.2.2 Unmatched Case, H_0

A randomly generated test patch with the same design parameter Γ as the reference patch, is almost surely to be different from the reference patch. In this case, the number of 0-to-1 flips δ_m does not equal to the number of 1-to-0 flips δ_n, hence the more general relationship described in Eq. (3.1) should be used. With the deterministic relationship between $\hat{\rho}$ and (δ_m, δ_n), we derive the joint distribution $f_{\Delta_m,\Delta_n|H_0}(\delta_m, \delta_n)$ for evaluating the distribution of $\hat{\rho}$ under H_0.

A randomly positioned test circle can have overlap with a reference circle only if the test circle is within $2r$ range of the reference circle as depicted in Fig. 3.16 (a). Then, the various subcases of matching a random test surface image with n circles to a reference image with same number of circles can be combined to dropping n points into $n+1$ holes: n of which are circular regions of radius $2r$, and the remaining of which is the surface container region with no $2r$ circular regions. We impose a few geometric constraints to ease the modeling procedure as illustrated in Fig. 3.16 (b): i) circles in the reference image are at least $3r$ away from the boundaries of surface containers to avoid the test circles being fully/partly outside the container when they overlap with a reference circle, and ii) circles in the reference image are at least $6r$ away from each other to avoid the overlap of two test circles when they overlap with two different reference circles respectively. Hence, the probability of overlapping with one circle is $p = \frac{\pi(2r)^2}{a^2}$, and the probability of not overlapping with any circle is $(1 - np)$, which can be modeled as a conceptual ball-and-urn problem, as shown in Fig. 3.16 (c).

It is intuitive to use multinomial distribution to model the current scenario of throwing n dots into $n+1$ holes with the following closed-form PMF:

$$\begin{aligned} p_{\mathbf{K},R}(\mathbf{k}, r) &= \frac{n!}{k_1! \cdots k_n!} p^{k_1} \cdots p^{k_n} (1 - np)^r, \\ & k_1 + \cdots + k_n + r = n \end{aligned} \tag{3.4}$$

where $\mathbf{k} = (k_1, \ldots, k_n)$, k_i is number of dots fallen in ith hole, and r is number of dots fallen outside all holes. Conditioned on a particular realization of the $(n+1)$-tuple $(k_1, \ldots, k_n; r)$, the joint distribution (Δ_m, Δ_n) can be approached analytically or numerically.

However, using the multinomial distribution as the prior does not lead the modeling too far. Different cases given by the multinomial prior can correspond to

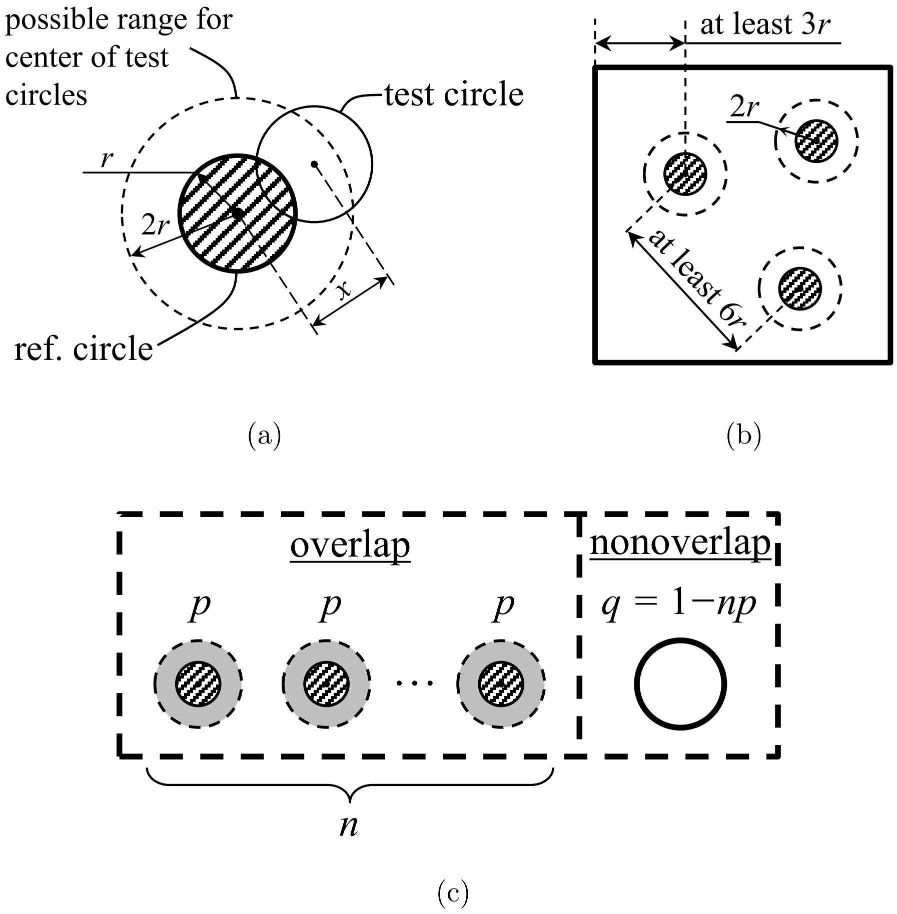


Figure 3.16: Auxiliary schematic for understanding how randomly positioned test circles overlaps with deterministically positioned reference circles: (a) possible range for center of test circles given the position of reference circle, (b) geometric constraints imposed on the locations of reference circles to ease the modeling procedure, and (c) circle matching problem simplified to a conceptual ball-and-urn problem.

a same case that we are interested in to analyze, *e.g.*, multinomial cases

$$(1,0,0,\cdots,0;n-1), \tag{3.5a}$$

$$(0,1,0,\cdots,0;n-1), \tag{3.5b}$$

$$(0,0,1,\cdots,0;n-1), \textit{etc.} \tag{3.5c}$$

correspond to a single analytical cases that there is only one circle in the test image overlap with one circle in the reference image, no matter on which reference circle the overlap happens. Hence, a better prior treating the above cases together should be constructed.

The problem is more tractable by focusing on cases that there are a number of test circles, k, overlapping with a most n reference circles, *i.e.*, $k_1 + \cdots + k_n = k$. Formally, we define an event $A_k = \{k$ circles from the test image overlap circles in the reference image, and $(n-k)$ circles do not.$\}$, and the PMF is

$$\mathbb{P}(A_k) = \binom{n}{k}(np)^k(1-np)^{n-k}, \qquad k = 0, \cdots, n. \tag{3.6}$$

The problem of k test circles overlapping with any number of n reference circles can be viewed as the classic problem of k points in n holes, *i.e.*,

$$k_1 + \cdots + k_n = k, \qquad 0 \leq k_i \leq k,\ i = 0, \cdots, n. \tag{3.7}$$

The total number of possible $(k_1, \ldots, k_n)$'s (formally called "compositions" in combinatorics) is well known to be $\binom{k+n-1}{n-1}$. We are interested in grouping all compositions with same elements together, and then calculating the probability of each group (formally called "partition"), since the partition is the largest unit that the joint distribution (Δ_m, Δ_n) takes the same form.

Table 3.1 enumerates all partitions for $k = 4$. The number of compositions in jth partition $B_j^{(k)}$, $k = 4$ are also shown, and the total number of composition is verified. The probability of each partition can be therefore calculated, and the PMF $\mathbb{P}(B_j^{(k)})$, $j = 1, \cdots, J(k)$ is obtained for $k = 4$ dots and n holes, in which $J(k)$ is the number of distinct symbols (partition) of the PMF. The function $J(k)$ does

Table 3.1: Possible partitions and number of compositions for each partition when there are $k = 4$ ball and n holes

Partition	Subgroup size	#Covered holes	#Compositions
$B_1^{(4)} = \{1, 3\}$	$1, 1$	2	$\binom{n}{1,1,n-2}$
$B_2^{(4)} = \{1, 1, 2\}$	$2, 1$	3	$\binom{n}{2,1,n-3}$
$B_3^{(4)} = \{1, 1, 1, 1\}$	4	4	$\binom{n}{4,n-4}$
$B_4^{(4)} = \{2, 2\}$	2	2	$\binom{n}{2,n-2}$
$B_5^{(4)} = \{4\}$	1	1	$\binom{n}{1,n-1}$
			total: $\binom{n+3}{n-1}$

not have an analytical form. It is called the *partition function* which increases fast as k increases.

When a surface uses n circles, PMFs for all $k = 0, \cdots, n$ are needed for analytical modeling. In our work, we use the best known integer partition algorithm to enumerate all partitions and then calculate all PMFs needed.

Under H_0, the joint distribution can be decomposed as

$$f_{\Delta_m,\Delta_n}(\delta_m, \delta_n) = \sum_{k=0}^{n} \sum_{j=1}^{J(k)} f_{\Delta_m,\Delta_n|B_j^{(k)},A_k}(\delta_m, \delta_n) \cdot \\ \mathbb{P}(B_j^{(k)}|A_k) \cdot \mathbb{P}(A_k) \tag{3.8}$$

where it is understood that all probability measures are under H_0, and $f_{\Delta_m,\Delta_n|B_j^{(k)},A_k}(\delta_m, \delta_n)$ describes the joint distribution of area of 0-to-1 flips and area of 1-to-0 flips when k test circles are thrown into holes of reference circles resulting an allocation pattern described by the jth partition.

The conditional joint distribution $f_{\Delta_m,\Delta_n|B_j^{(k)},A_k}(\delta_m,\delta_n)$ is analytically intractable for $k \geq 3$. However, simulation for a partition $B_j^{(k)}$ for each pair of (j,k) is doable and takes much less memory and time than simulation for a whole patch.

Take $B_2^{(4)}$ listed in Table 3.1 as an example. Four balls are put into three "covered holes" with 1, 1, 2 balls for hole 1, hole 2, and hole 3, respectively. We denoted the number of covered holes by #ch. Then, the size of two types of flipped regions can be aggregated from the contributions from each covered hole:

$$\begin{cases} \delta_m = \sum_{\ell=1}^{\#\mathrm{ch}} \delta_m^{(\ell)} + (n-k)\pi r^2, \\ \delta_n = \sum_{\ell=1}^{\#\mathrm{ch}} \delta_n^{(\ell)} + (n-\#\mathrm{ch})\pi r^2. \end{cases} \tag{3.9}$$

Using simulation, we can easily obtain a large number of the sets $\{(\delta_m^{(\ell)}, \delta_n^{(\ell)})\}_{\ell=1}^{\#\mathrm{ch}}$, say, 1000. We aggregate them using Eq. (3.9) and obtain 1000 pairs of (δ_m, δ_n), which can be considered as a reasonably large sample drawn from the population $f_{\Delta_m,\Delta_n|B_2^{(4)},A_4}(\delta_m,\delta_n)$. Passing the sample through the deterministic correlation function of Eq. (3.1) leads to a sample of size 1000 of the random variable $\hat{\rho}$. Proper methods for estimating the distribution for $\hat{\rho}$ can be subsequently applied.

3.6.3 Model Validation

The model minimizes the use of simulation by decomposing the big simulation problem into three levels, and simulates at the lowest level only. It has the advantage of using less memory and low computational complexity.

In this section, we compare the PDF under H_0 obtained via modeling and via simulation. As the mean correlation under H_0 is zero, we compare the standard

Table 3.2: Standard deviation of PDF under H_0 evaluated from model vs. from simulated patches (patch size: 200)

Radius	1	2	3	4	5	6	7	8	9	10
Simulated	0.008	0.014	0.020	0.026	0.032	0.038	0.043	0.050	0.056	0.062
Model	0.006	0.012	0.018	0.024	0.029	0.035	0.041	0.046	0.051	0.055

deviation obtained from the two methods. Table 3.2 reveals that the modeling results are close to the simulated results that serve as the anchor. This confirms the correctness of the proposed generative model.

The proposed generative model allows us to examine the authentication performance under different choices of parameters, such as the shape, size, and number of foreground objects. The findings will lead to a systematic understanding of the class of extrinsic paper surfaces, and will facilitate the design of surfaces in practical counterfeit detection scenarios.

4

Invisible Geo-Location Signature in a Single Image

4.1 Chapter Introduction

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

A number of new generations of digital cameras today, including many cell-phone cameras, are capable of associating the built-in GPS input with the images and videos produced by the camera in the metadata field regarding the location in which these images/videos were captured. These metadata, however, are not always available, and they can be tampered with in a relatively easy way.

Recent advances, notably through the FINDER program by the U.S. Intelligence Advanced Research Project Activity (IARPA) agency [26], utilized computer vision techniques and a big data paradigm to exploit visible terrains and landmarks appearing in an image and match them with a variety of geospatial data to identify the image's geographic location. These technologies would still face considerable challenges in cases of absence of visible landmarks, such as when the background of outdoor images does not have differentiating features, or when an image is captured indoors. The latter scenarios are common in such global efforts as fighting crime on child exploitations [27–29], where a significant portion of images of child pornography and the related abuse have been taken indoors in private locations by perpetrators, and the Internet has eliminated the national borders on such crime and increased the difficulty in tracking down the perpetrators. Having new technologies to locate where such images were taken, even on a coarse level, will provide an unprecedented tool to aid the global investigation, allowing law enforcement at appropriate jurisdictions.

In this chapter, we explore a nearly invisible signature about location that is inherently captured in an image. This invisible signature comes from the power distribution network whose varying frequency properties over time are known as the Electric Network Frequency (ENF) signal. There is a growing amount of multimedia forensics research recently on the ENF signal, as the ENF signature over time in a multimedia file reflects the behavior of the power grid at the time and the location where the media signal is recorded. Such a relation enabled the estimation of the time, location and integrity of the corresponding multimedia signals [30–34].

Due to the temporally varying nature of the ENF signature that plays a critical role in its ability to relate with time and location, ENF extraction in the literature thus far requires the hosting signal to have the temporal nature as well, for example, an audio or a video. Two intriguing questions to push the boundary of ENFs localization capability are: Can ENF traces be found in a single image? And can such a basic attribute as the nominal ENF value be extracted at a reasonable accuracy from a single image? This chapter aims at answering these questions where by narrowing down whether an image under question comes from a 50Hz or 60Hz country would provide a valuable clue to law enforcement investigations and shine a light on the jurisdiction in charge. To the best of our knowledge, this is the first work toward developing a unique machine vision capability to address the challenging geo-location scenarios for individual images that do not have a visible landmark or GPS tag, especially those captured indoors. Given the devastatingly high estimates by the U.S. and United Nation authorities that over 300,000 U.S. children are sexually exploited every year and about a million children every year are forced into prostitution in some form worldwide [27], the societal impact of any new technology breakthrough from computer vision would be tremendous. The investigations presented in this chapter form an important first step toward this.

The rest of this chapter is organized as follows. Chapter 4.2 reviews the background information on the ENF signature and the image capturing operations that allow it to be present in images of interest. Chapter 4.3 presents our proposed approaches for extracting the ENF signature from a single image. Chapter 4.4 shows the experiments we conducted and the results obtained. Chapter 4.5 provides further discussions and Chapter 4.6 concludes the chapter.

4.2 Background and Preliminaries

4.2.1 Electric Network Frequency

The Electric Network Frequency (ENF) of power distribution networks has a nominal value of 60Hz in most of the Americas, and 50Hz in most of the rest of the world. The instantaneous frequency of the sinusoidal variation in power networks does not stay at this nominal value. Rather, it fluctuates around the nominal value due to load changes across the power grid. The trends in the ENF variations tend to be very similar at different points in the same grid. The changing instantaneous value of the ENF over time is what we define as the *ENF signal*.

The ENF signal has been a growing subject of multimedia forensics research in recent years due to its presence in media sensing signals. If an audio or video recording is made in an area where there is electrical activity, it is likely that this recording will capture the ENF variations at the time of recording. In audio recordings, this has been attributed to the acoustic mains hum emitted from devices connected to the power mains [35]. In video recordings, this comes from the near-invisible flickering of electric lighting [3].

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

4.2.2 Rolling Shutter and Read-Out Time

Unlike cameras that employ a *global shutter* that acquires the pixels of an image frame all at the same time, a camera employing rolling shutter acquires an image frame one row at a time. Although traditionally regarded as the cause of negative artifacts in images/videos, the rolling shutter has been exploited in beneficial ways for computational photography applications [36, 37]. In this chapter, we shall see that this sequential read-out time mechanism of the rolling shutter while capturing a single image allows the resulting image to capture samples of an incoming electric light signal at slightly different points in time, and thus obtain a short segment of ENF traces.

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

In order to extract ENF information found in a single image, it is important to know the read-out time T_{ro} value of the camera that had produced the image under question. In our work, we have made use of a protocol inspired by the one

Table 4.1: Estimated parameters for rolling-shutter images.

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

proposed in [38] to compute the T_{ro} value of a camera at hand. Table 4.1 lists the estimated values we computed for the cameras on the backsides of different models of iPhone devices that are used in this chapter.

4.2.3 ENF Signal Embedding in a Single Image

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

ENF-containing light signal as follows:

$$e[i] = \cos\left(2\pi f_{\text{ENF}} \cdot (iT_{\text{row}}) + \phi_0\right) \tag{4.1a}$$

$$= \cos\left(2\pi \frac{f_{\text{ENF}}}{f_{\text{row}}} \cdot i + \phi_0\right), \tag{4.1b}$$

where f_{ENF} is the fluctuation frequency of the light intensity, j is the row index of the image ranges from 0 to $L-1$, T_{row} is the portion of the read-out time assigned to each row, and ϕ_0 is the initial phase of ENF signal at time the first row is acquired.

When an image is taken through the rolling shutter mechanism, the light intensity signal is modulated with a native image $m_0[i,j]$. Although this modulation can be nonlinear, but a large part can be considered as an additive component $e[j]$ to the native image content. Hence, the ENF-containing image can be approximately modeled by adding a parametric surface to the native image content as follow:

$$m[i,j] = m_0[i,j] + e[j], \qquad \begin{aligned} i &= 1, \cdots, L, \\ j &= 1, \cdots, W, \end{aligned} \tag{4.2}$$

where m is the ENF containing image, and W is the number of columns of the image.

4.3 Proposed Framework and Approaches

4.3.1 Formulation and Framework

In this section, we address the following research question: *given an image with an inherently captured ENF signal that is generally very weak, is it possible to determine whether the nominal ENF value f of the embedded ENF signal is 100Hz or 120Hz?*

It is helpful to note that the instantaneous ENF frequency during a stable operation of most well managed grid has only a small deviation from the nominal value: typically within ±0.5Hz, and generally well within ±0.1Hz in the U.S. grids. An intuitive way to address the research question would be to formulate this as a general parameter estimation problem for the frequency parameter f. After obtaining the estimated frequency in the unit of Hz, one can then make a binary decision between 100Hz and 120Hz based on distances from the estimate to the two nominal values. Alternatively, one may bypass an explicit frequency estimation from analyzing the image and carry out a more direct classification by enforcing strong prior knowledge. The challenge for either approach is that the ENF traces can be very weak, nearly invisible. As such, developing a framework to tackle this problem would benefit from gaining insights on how this problem would be approached from a human vision point of view.

We have created a dataset to facilitate the research. We use multiple smartphone cameras to take images in primarily indoor environments from countries with nominal ENF values of 60Hz and 50Hz, respectively. For each scene, different levels of *exposure bias* were set on the camera using each camera's built-in functionality to control the exposure time of the captured images. A less negative exposure bias corresponds to a longer exposure time and a brighter image, and vice versa. Fig. 4.1 shows four scenes from this ENF image dataset taken in Country A where the nominal ENF value is 60Hz, or equivalently, 120Hz for the brightness variations of electric lighting; for each scene, four levels of exposure bias were used and the resulting image shown. As we can see, a short exposure time can quite clearly cap-

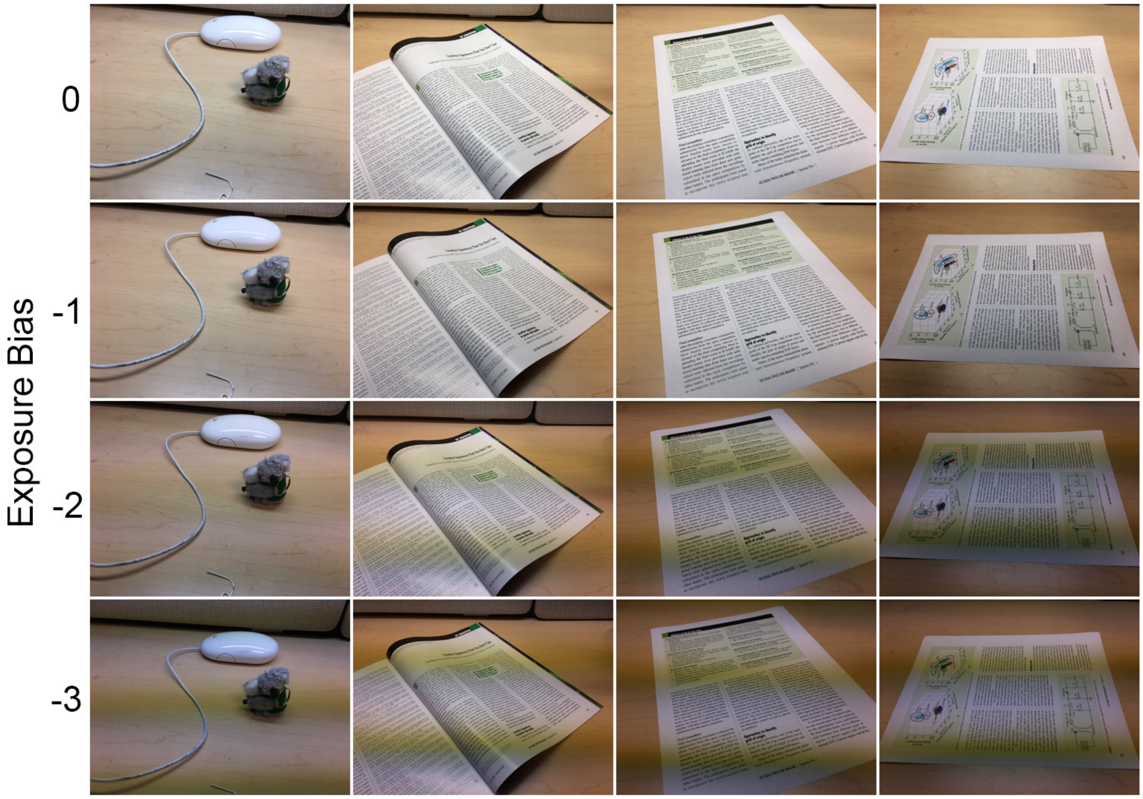


Figure 4.1: Four scenes from an images dataset took in Country A with ENF frequency at 60Hz using iPhone 6s. Each column corresponds to a different scene. From the top row to the bottom row, the exposure biases are set to 0, -1, -2, and -3, respectively, when taking the images. The more negative the exposure bias is, the stronger the ENF traces can be observed.

ture the fluctuation of the ENF light signal. As the exposure time increases, the amplitude of the fluctuation of the captured signal decreases due to an effective convolution of a rectangular window with the electric light signal.

If the task of determining the nominal ENF values would be given to a human observer, he/she is capable of using visible clues and can thus guess the parameters of the sinusoid surface fluctuating along the rows, namely, the frequency f, the magnitude A, and the initial phase ϕ_0. It is understandable that the guess is easier and can be more accurate for those images with stronger ENF traces, such as those from the two lower rows of Fig. 4.1.

One of the key challenge to extract an ENF from an image and estimate or classify the nominal ENF value is to differentiate the highly structured sinusoid pattern from its combination with the image content. We can draw insights from the intuitions of a human observer to inspire the design of effective computer vision algorithms.

First, a human observer would understand that the sinusoid induced by the power source is rather distinctive and independent from the native image content. By exploiting the regularity of the sinusoid signal as well as that of the native image content, one may obtain a reasonable separation. We take advantage of this insight of signal independence through the ICA-based method proposed in Chapter 4.3.2.

Second, a human observer may detect ENF traces by exploiting the recurring sinusoid patterns shown in different, even disconnected regions of an image. This insight of recurring sinusoid patterns is formalized by the parametric model proposed in Chapter 4.3.3.

Third, a human's confidence of a good separation is dependent on the structure of the native image content. The extraction of the sinusoid signal is easier over smooth regions than over "busy" regions such as those with textures. Even if the sinusoid signal spans across two smooth regions separated by an edge, the gradual change due to the sinusoidal ENF signal in the individually smooth regions can still reveal a good amount of clue as compared to the case of the sinusoid signal being added to a smooth region. The entropy-based method proposed in Chapter 4.3.3 leverages this insight.

Meanwhile, computer vision has the advantage over human vision in discovering weak signals and changes that are below the visibility threshold of human observers. By employing suitable mathematical tools, we have the potential to overcome the limitation of human vision, and therefore push forward the limit of the identification capability in terms of pixel intensity. Next, we present the two algorithms to tackle the ENF classification problem from image.

4.3.2 Independent Component Analysis Based Signal Separation

The independent component analysis (ICA) is a power *source separation* tool to recover independent random variables from a collection of observed random variables that are linear combinations of them [39]. Denote a collection of three independent random variables as $\mathbf{s} = [s_1, s_2, s_3]^T$, and the operations of the linear combinations by a 3-by-3 nondegenerated *mixing* matrix $\mathbf{M}$, the generative process for the observed random variables $\mathbf{x} = [x_1, x_2, x_3]^T$ is defined as

$$\mathbf{x} = \mathbf{M}\,\mathbf{s}. \tag{4.3}$$

The ICA seeks to simultaneously solve for the independent random variables $\mathbf{s}$ and the mixing matrix $\mathbf{M}$. It is equivalent to writing the estimate $\hat{\mathbf{s}}$ in terms of $\mathbf{x}$ and $\mathbf{M}$ and finding best $\hat{\mathbf{s}}$ and $\mathbf{M}$ such that the joint distribution of $\hat{\mathbf{s}}$ factors into marginal distributions, i.e.,

$$p(\hat{s}_1, \hat{s}_2, \hat{s}_3) = p(\hat{s}_1) \cdot p(\hat{s}_2) \cdot p(\hat{s}_3), \tag{4.4}$$

where $p(\cdot)$'s are corresponding probability distributions. The solution of the ICA usually does not have a closed form, so one has to resort to optimization. Objective functions that can quantitatively measure the degree of independence among $\hat{s}_i$'s is needed. One such function, deeply rooted in the information theory [40], is the Kullback–Leibler (KL) divergence. The "distance" from $p(\hat{s}_1, \hat{s}_2, \hat{s}_3)$ to $\prod_{i=1}^{3} p(\hat{s}_i)$ can be mapped to the range of $[0, 1]$ to quantify the degree of independence.

In the problem of separating the sinusoid signal from the image, the independence of the foreign sinusoid signal and the native image content allows the direct application of the ICA. Consider one column from the image as an example. The column of image data is composed of three vectors corresponding to three different color channels. They are considered as the realizations for three observed random variables, in which each row is one realization. After a 3-by-3 demixing process, realizations of three underlying independent random variables are obtained. One of them should correspond to the native image content, and another should correspond to the sinusoid signal. Frequency estimation can be applied to the sinusoid component to obtain a frequency estimate for deciding between 100Hz and 120Hz.

We can repeatedly apply the ICA to D different columns of the image, and then fuse the frequency estimates from different columns into the final one. However, the resulting recovered components for different columns are not aligned. For example, sinusoid signals may be found in the first recovered component of dth column, $\hat{s}_1^{(d)}$, and in the second recovered component of $(d+1)$st column, $\hat{s}_2^{(d+1)}$. This inconsistency in the index of the recovered component can be addressed by the independent vector analysis (IVA) [41], an extension of ICA by considering correlations among different groups, or columns, in our example. The key change in the modeling assumption is to replace each random variable appeared in Eq. (4.4) by the collection of correlated random variables from D groups, i.e., $\hat{s}_i \leftarrow \hat{\mathbf{s}}_i = [\hat{s}_i^{(1)}, \hat{s}_i^{(2)}, \cdots, \hat{s}_i^{(D)}]^T,\ i = 1, 2, 3$. The pseudo code for the proposed ICA-based method is shown in Algorithm 1.

4.3.3 Entropy Minimization for Parametric-Surface–Removed Images

The ICA-based method does not fully utilize the prior knowledge that one of the two signals to be separated is a sinusoid signal, which can potentially limit its performance. In this section, we propose a method that explicitly exploits the parametric property of the ENF signal, and that exploits the change of a statistical property before and after ENF embedding.

One observation for constant signals corrupted by sinusoid signals is that, their histograms after corruption become more spread. That is, the sole bin of the histogram before corruption starts to split into multiple bins due to the positive and negative additive values contributed by the sinusoid signal. This spreadness of the histogram corresponds to the randomness of a random source/variable that can be constructed from the histogram, and one measure for such randomness is the entropy. Intuitively, before the corruption, the histogram is an impulse and therefore with 0 entropy; after the corruption, the histogram spreads and therefore with an entropy greater than 0. It is straightforward to establish a rigorous mathematical result that adding a real-valued sinusoid signal with uniformly random initial phase will increase the entropy of a constant signal. Simulation result in Fig. 4.2(a) reveals that 99% of all columns of an image have an increase in entropy after being corrupted by sinusoid signals. From an image compression point of view, a constant region is much easier to encode and therefore has lower entropy than a sinusoid-corrupted region.

With the result of entropy change due to addictive sinusoid signal, we define an objective function as the average of the entropy of columns of the sinusoid-surface–removed image with three parameters: frequency f, amplitude A, and initial phase ϕ_0. Ideally, if the task is to estimate f, one needs the objective function to be jointly convex over (f, A, ϕ_0). However, the task is to differentiate images captured in 50Hz countries from those captured in 60Hz countries, and therefore only two possible choices of f exist. Incorporating this special structure of the problem can relax the stringent joint convexity constraint to require the objective function merely convex over (A, ϕ_0). Fig. 4.2(b) of error surfaces for 25 columns of an image supports that, when the frequency is a correct guess, the objective function are generally convex in (A, ϕ_0).

We now verify if the frequency decision task can be solved with the two properties mentioned above. Given an image corrupted by one of the two possible ENF signals. We estimate (A, ϕ_0) separately for $f = 100$Hz and $f = 120$Hz. When the estimation is carried out for the correct frequency, the residue is roughly the native image content per the convexity property; when the estimation is carried out for incorrect frequency, the residue is a combination of the native image content, and the sum of two sinusoid signals with different frequencies. Given that the native image content has lower entropy than the content with additional sinusoid components, we choose the candidate frequency leading to the lower entropy as our decided frequency. The pseudo code for the proposed entropy minimization method is shown in Algorithm 2.

4.4 Experimental Results

In this section, we show the performances of our proposed methods. We first test the two methods using synthetic data, which reveals that the entropy minimization method provides better differentiation of the nominal ENF value than the ICA-based method at weak ENF amplitudes. We then carry out and analyze an in-depth experiment on the entropy minimization method with a real-world image dataset.

4.4.1 Results of Independent Component Analysis Based Method

We use six images that do not contain ENF traces as native images to synthesize test images captured in 50Hz and 60Hz regions. Sinusoid signals with random initial phases and a list of decreasing amplitude levels, $[16, 14, 12, 10, 8, 4, 2, 1, 0.5]$ in the same unit of pixel intensity with 256-level shades of gray were added to the native images to generate a total of 108 $(= 6 \times 2 \times 9)$ images. For each image, the proposed algorithm of stochastic nature was applied ten times to better reveal performance of the algorithm on the particular image. Performances for 50Hz and 60Hz images were assessed separately.

The estimated probability of success decision, $\hat{p}$ (*aka* the estimated success rate), and its standard deviation quantifying the confidence of the estimate were shown in Fig. 4.3 as a function of decreasing ENF amplitude. The curve for 50Hz images is always above 0.8 success rate, whereas for 60Hz images the performance drops to random guesses when the amplitude is reduced to around 5. Taking images with both frequency values into consideration, we conclude that the proposed method performs better than random guesses when the amplitude the sinusoid signal is greater than 5.

As we examine the intermediate data of the method for the low amplitude cases, we can see that the peaks of the power spectrum are often biased toward the low frequency. This is because when the amplitude of the ENF signal decreases, the power spectrum becomes increasingly dominated by the image contents that contain slowly-changing, low-frequency signals. Therefore, at low ENF amplitude, the frequency estimates by the ICA-based method are intrinsically biased toward low frequency. Hence, the success rate curve for the candidate of the higher frequency, i.e., 60Hz, better reflects the true performance of the proposed method.

4.4.2 Results of Entropy Minimization Method

4.4.2.1 Images with Synthetic ENF Traces

We use the six images as in Chapter 4.4.1 to synthesize test images captured in 50Hz and 60Hz regions. Sinusoid signals with random initial phases and a list of decreasing amplitude levels, $[16, 8, 4, 2, 1, 0.5]$, were added to the native images to generate a total of 72 ($= 6 \times 2 \times 6$) images. For each image and each tested number of columns used, the proposed algorithm of stochastic nature was applied five times.

The estimated success rate as a function of decreasing ENF amplitude under the different values of number of columns used (*aka* #cols) were shown in Fig. 4.4 for 50Hz and 60Hz images, respectively. It is revealed that the success rate for both cases can be as high as 0.9 for ENF amplitude equals to 4. As the amplitude decreases further, the success rate drops to random guesses when amplitude is around 0.5 for 50Hz images, and is between 0.5 and 1 for 60Hz images.

Examining the accuracy of the estimators for the nuisance parameters A and ϕ_0 can hint on the reliability of the proposed entropy minimization method. For example, Fig. 4.5 shows that as the number of columns used increases, the mean absolute error (MAE) generally decreases. Interestingly, both plots for 50Hz and 60Hz images show that increasing from one column to two columns leads to the largest improvement in accuracy, and further increase in number of columns used leads to marginal improvement.

4.4.2.2 Real Rolling Shutter Images with ENF Traces

When collecting real-world camera images, we do not have a direct control on the amplitude of the embedded ENF signal. However, we can indirectly affect the amplitude by changing the *exposure bias* parameter related to the exposure time of each row. Fig. 4.1 shows four scenes captured under 120Hz lighting with different exposure bias values. The amplitude of ENF with the same exposure bias across different scenes does not seem to vary significantly. We acquired in Country A a dataset with 60Hz ENF signal containing six scenes, each scene with exposure bias of $[0, -1, -2, -3]$; and in Country B a dataset with 50Hz ENF signal containing five scenes, each scene with exposure bias of $[0, -1.5, -1.8, -2.1, -2.4, -3]$.

Fig. 4.6 shows the performance of the proposed entropy minimization on real-world camera images captured in a 60Hz ENF country. One can observe that compared to using two columns to obtain a strong detection performance as in the synthetic case of Chapter 4.4.2.1, 16 columns are needed in the real-world camera image case. It is also revealed that as the level of exposure bias reduces from -2 to -1, the method breaks as the success rate becomes dominated by being biased toward deciding 50Hz. Fig. 4.7 shows the estimates for the nuisance parameters (A, ϕ_0). It is revealed that the range of amplitude under exposure bias of -2 is from 10 to 20, and that under exposure bias of -1 is below 10. Combining the two aspects of the results, we anticipate that the current algorithm may start to fail on real-world camera images when the amplitude is below 10.

Fig. 4.8 for real-world camera images captured in a 50Hz country similarly shows that as the number of columns used increase, the success rate and estimation accuracy improves. One interesting observation, as has been discussed when being observed in Fig. 4.3, is that no matter how weak the ENF signal is, the success rate for 50Hz images is never below random guesses. This suggests that even though the proposed entropy minimization method does not explicitly estimate the frequency that results in biasing toward low frequency, and the experimental results for images with synthetic ENF traces do not show such bias, the bias does exist when the entropy minimization method is applied on the real-world camera images. Further investigations can be carried out to examine the impact of the modeling mismatch, e.g., when the ENF signal captured in the image is not a perfect sinusoid.

4.5 Discussions

It is important to recall the significant challenge in extracting ENF traces from a single image, as it is extremely difficult to overcome the dominating visual content and reveal the weak ENF traces from just one image. In contrast, the ENF extraction from video [1,3] can take advantage of multiple frames to estimate the native visual content and separate the ENF from the visual content. Despite the difficulty, the results in Chapter 4.4 have shown promising capability by our proposed algorithms as the first attempt to tackle the challenge.

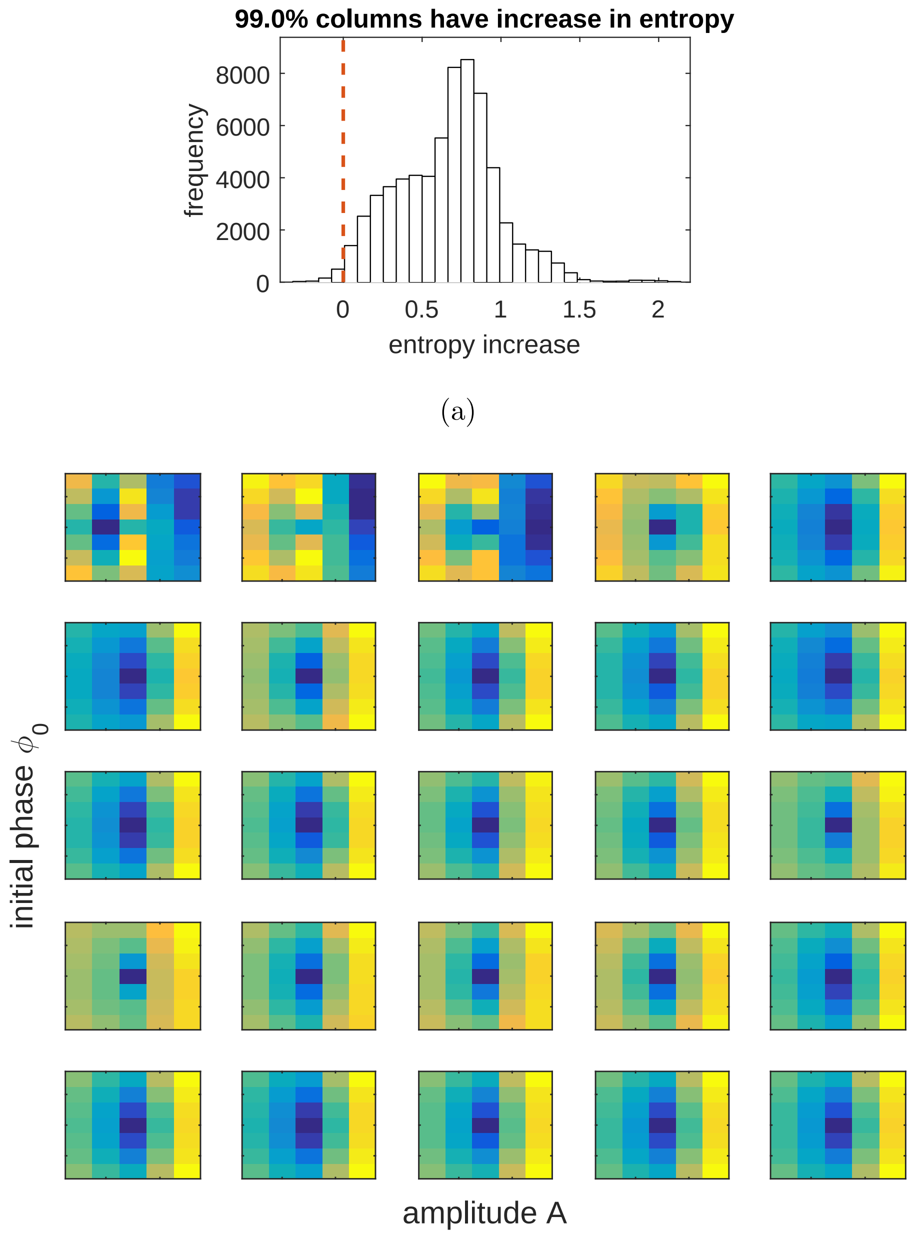
The ICA and other source separation methods have been widely used in biomedical signal processing applications for recovering independent signals from mixed observations. In recent work on heart-rate signal extraction, ICA and other source separation methods have been found effective extracting sinusoid signal-like heart-rate signal from the periodic change of color in human faces [42, 43]. Our proposed ICA-based decision method was partly inspired from such successes. As the experimental results have shown, its performance is not as good as the entropy minimization method that is specially designed from the problem of ENF signal decision by making use the parametric signal form. Furthermore, the ICA-based method assumes probabilistic sources and i.i.d. realizations, which is not true for the ENF signal as the sinusoid is deterministic and sample points are dependent. This violation of modeling assumption can contribute to its limited capability of signal separation, as hinted by the experimental results.

4.6 Chapter Summary

In this chapter, we explored the use of an invisible power signature, the ENF signal, that may be embedded in images taken by CMOS cameras for geo-tagging purposes. Specifically, we addressed the research question of whether the basic attributes of an embedded ENF signal can be correctly identified. We proposed two algorithms, the ICA-based method and the entropy minimization method. Experimental results show that both methods are able to make high accuracy decisions when the ENF traces are strong, whereas the entropy minimization method outperforms the ICA-based method as ENF traces become weaker. In this proof-of-concept work, we have demonstrated a unique computer vision capability of extracting invisible traces to help narrow down the capturing geographic region of an image. The next step would be to examine and improve the performance on a larger scale investigation with more cameras and images, and explore additional location information that can be inferred from the ENF traces in an image.

Algorithm 1: The proposed ICA-based decision method for images with embedded ENF traces.

```
input  : An RGB image with embedded an ENF signal at $f_{\text{ENF}} \in \{100, 120\}$ Hz
output: $\hat{f}_{\text{ENF}} \in \{100, 120\}$ Hz
Step 1. Initialization
Randomly select $D$ columns, each comprised of three color channels, as the
  realizations for observed random variables $\{x_c^{(1)}\}_{c=1}^3, \cdots, \{x_c^{(D)}\}_{c=1}^3$
Step 2. Source Separation
Apply IVA to obtain realizations of independent random variables
  $\{s_c^{(1)}\}_{c=1}^3, \cdots, \{s_c^{(D)}\}_{c=1}^3$
Step 3. Frequency Estimation
for $c \leftarrow 1$ to 3 do
    for $d \leftarrow 1$ to $D$ do
        $\mathbf{s} \leftarrow$ realizations of $s_c^{(d)}$
        $[\hat{\mathbf{F}}(c, d), \hat{\mathbf{P}}(c, d)] \leftarrow$ PeakFreq&Spec(s)
    end
    $\hat{\mu}_f(c) \leftarrow$ RobustMean($\hat{\mathbf{F}}(c, :), \hat{\mathbf{P}}(c, :)$)
end
Step 4. Decision
Decide $\hat{f}_{\text{ENF}}$ between 100Hz and 120Hz based on the closest distance of $\hat{\mu}_f$ to these
  two values.
```



(a)

(b)

Figure 4.2: (a) Histogram for entropy increase for all columns of an image with embedded ENF signals at 50 and 60 Hz with ten random phases. (b) 2-D sample error surfaces as a function of amplitude A and initial phase ϕ_0 for 25 equally-spaced columns of an image. Legend: More bluish or darker colors mean lower error. (Figure is best viewed in color.)

Algorithm 2: The proposed entropy minimization decision method for images with embedded ENF traces.

```
input  : An RGB image with embedded an ENF signal at $f_{\text{ENF}} \in \{100, 120\}$ Hz
output: $\hat{f}_{\text{ENF}} \in \{100, 120\}$ Hz
Step 1. Initialization
Convert the image to gray-scale
Remove the trend of overall intensity by RANSAC-based detrending
Randomly select $D$ gray-level columns, as the realizations for observed random
  variables: $x^{(1)}, \cdots, x^{(D)}$
Step 2. Optimization
$f_{\text{arr}} \leftarrow [100, 120]$
$\mathbf{X}(:, d) \leftarrow$ realizations of $x^{(d)}, \quad \forall d$
for $k \leftarrow 1$ to $2$ do
    $br(k) \leftarrow \min_{A,\phi_0} \mathit{objFunc}\left(\mathbf{X}, f_{\text{arr}}(k), A, \phi_0\right)$
end
Step 3. Decision
Deicide $\hat{f}_{\text{ENF}}$ between 100Hz and 120Hz based on the one gives a smaller bitrate $br$
Subroutine $\mathit{bitrate} \leftarrow \mathit{objFunc}\left(\mathbf{X}, f, A, \phi_0\right)$
    $e(i) \leftarrow A\cos(2\pi \frac{f}{f_{\text{row}}} \cdot i + \phi_0), \quad \forall i$
    $\hat{\mathbf{X}}(i, j) \leftarrow \mathbf{X}(i, j) - e(i), \quad \forall (i, j)$
    $\mathit{bitrate}_{\text{arr}}(j) \leftarrow \mathit{calcEntropy}(\hat{\mathbf{X}}(:, j)), \quad \forall j$
    return $\mathit{mean}(\mathit{bitrate}_{\text{arr}})$
```

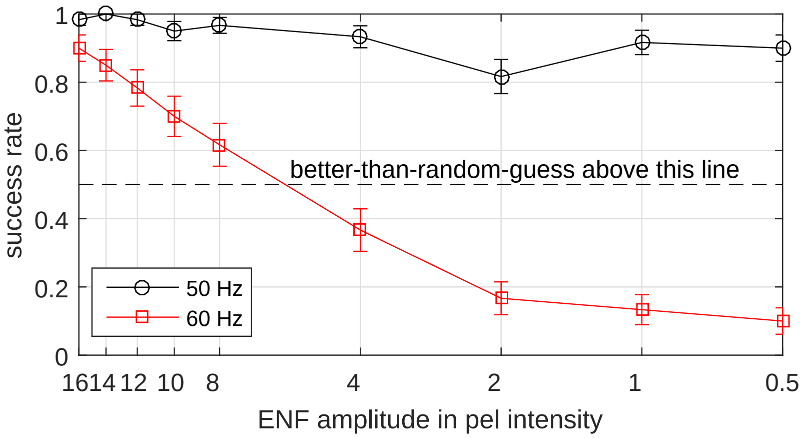
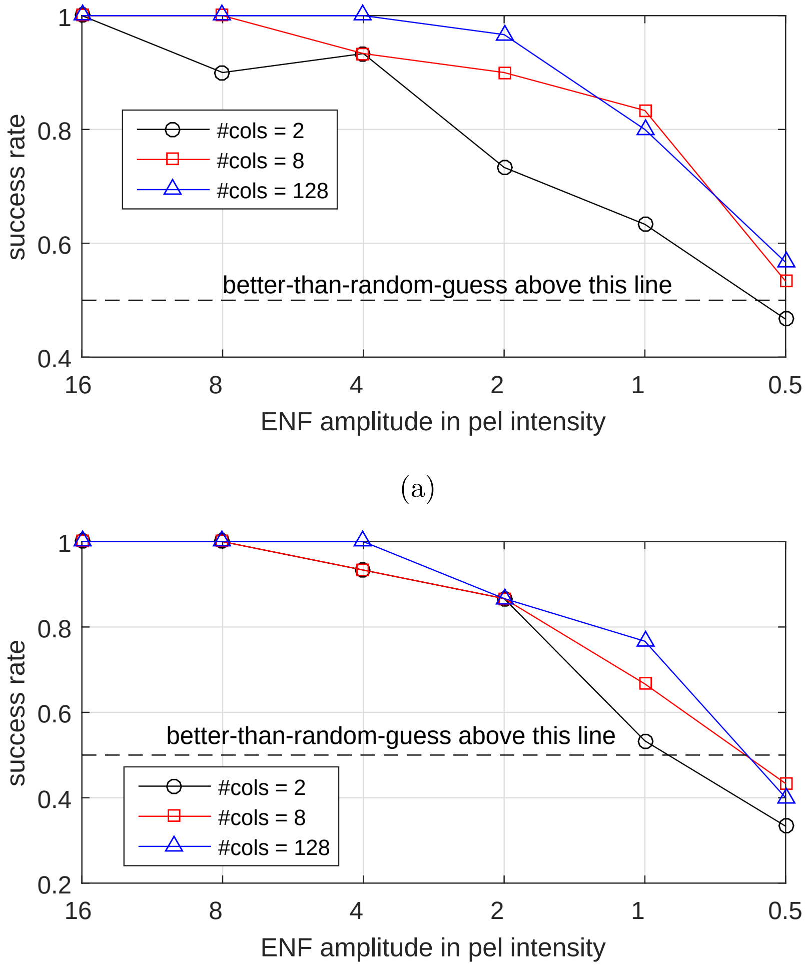


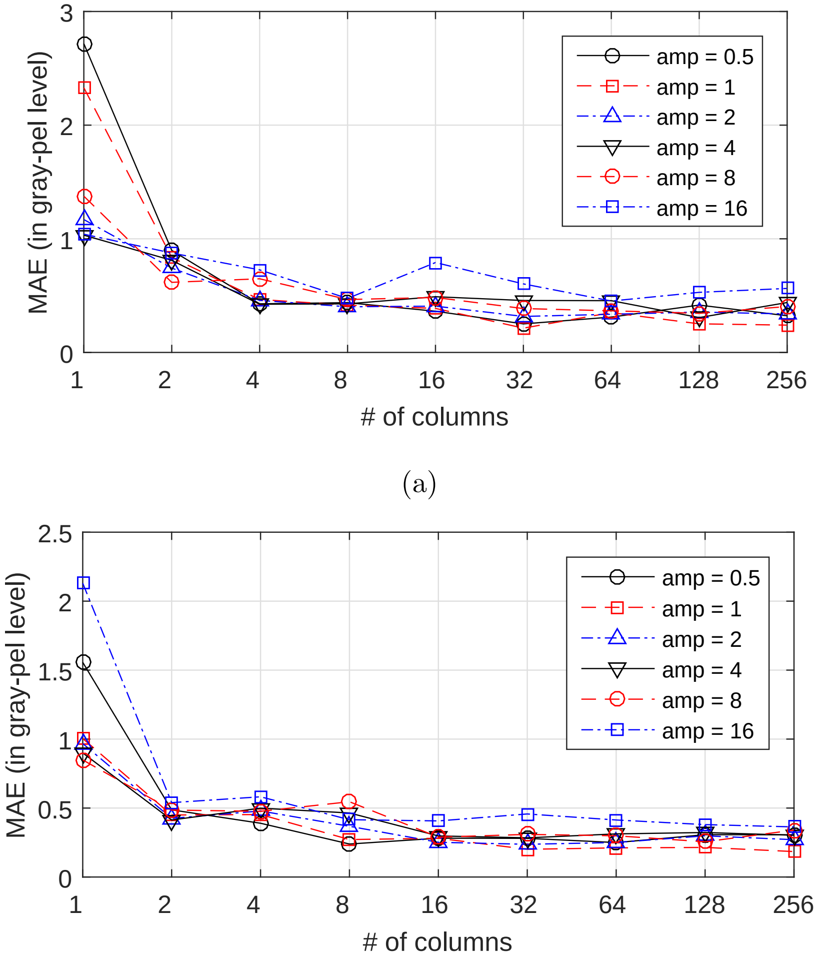
Figure 4.3: Success rate of using joint blind source separation to decide embedded ENF signal at 50Hz and 60Hz. As the amplitude of the synthetic ENF reduces, decisions for images with 60Hz ENF become worse; decisions for images with 50Hz ENF does not drop significantly due to the bias contributed by low spatial frequency image contents.



(a)

(b)

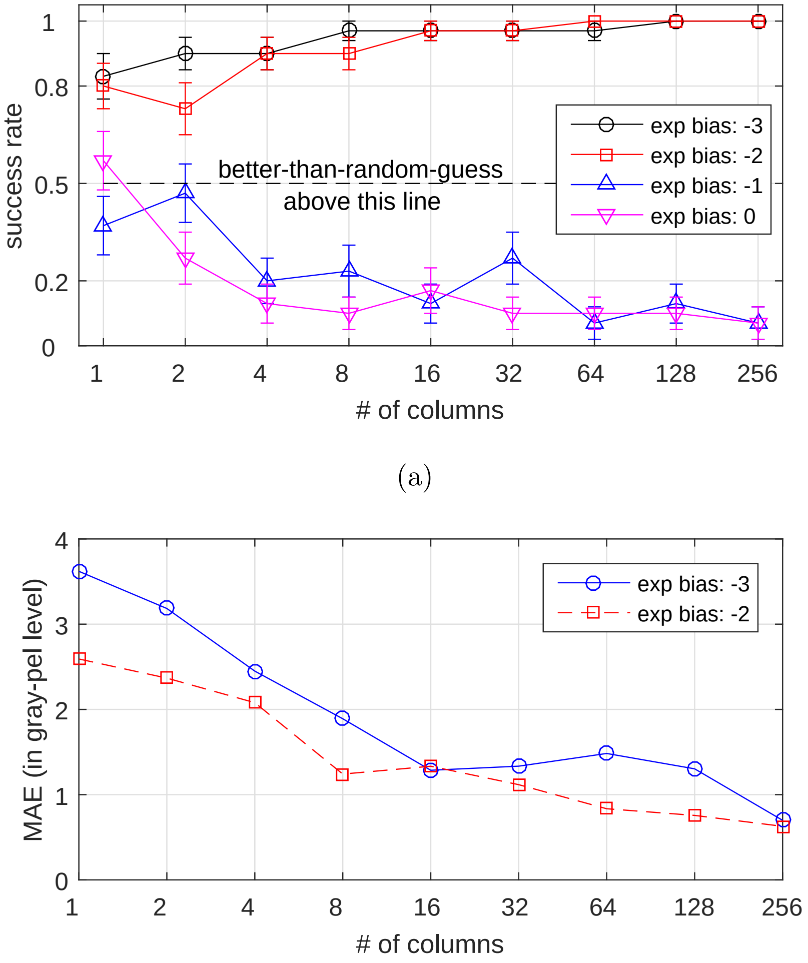
Figure 4.4: Success rate for deciding images with (a) 50Hz ENF traces, and (b) 60Hz ENF as a function of the ENF amplitude, and the number of columns used.



(a)

(b)

Figure 4.5: MAE for the amplitude estimator for (a) 50Hz images, and (b) 60Hz images.



(a)

(b)

Figure 4.6: (a) Success rate for deciding images with 60Hz ENF traces as a function of the ENF strength and the number of columns used in constructing the objective function. (b) MAE for the amplitude estimator.

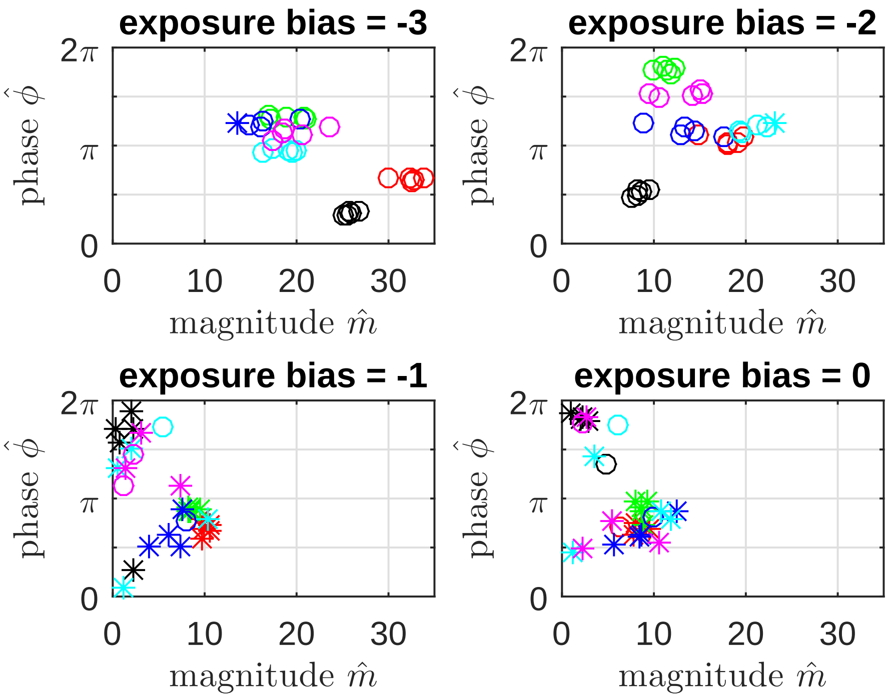
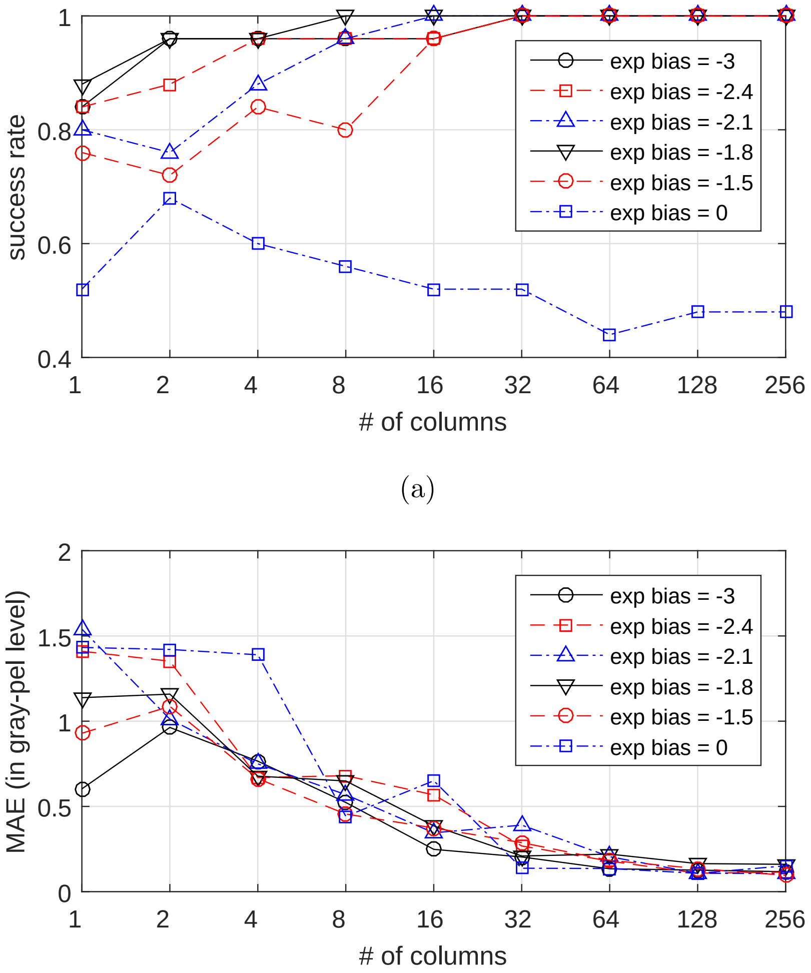


Figure 4.7: Estimates of nuisance parameters for images (one color per image) captured in Country A (60Hz) with different levels of exposure bias. Circles "○" and star "*" correspond to an estimate leading to a correct and incorrect decision, respectively.



(a)

(b)

Figure 4.8: (a) Success rate for deciding images with 50Hz ENF traces as a function of the ENF strength and the number of columns used in constructing the objective function. (b) MAE for the amplitude estimator.

5

Fitness Heart Rate Measurement Using Face Videos

5.1 Chapter Introduction

Contact-free monitoring of the heart rate using videos of human faces is a user-friendly approach compared to conventional contact based ones such as electrodes, chest belts, and finger clips. Such monitoring system extracts from a face video a 1-D sinusoid-like face color signal that has the same frequency as the heartbeat. The ability to measure heart rate without touch-based sensors is attractive and gives it potentials in such applications as smart health and sports medicine.

Heart rate from videos was first demonstrated feasible in [44], and since then most work [45–55] has been focusing on still/rest cases or those with relatively small body motions. In contrast, less work [56–58] has been on large motion scenarios such as fitness exercises. In [56], the authors did a proof-of-concept study showing that after using block-based motion estimation for a cycling exercise video, a periodic signal can be extracted from the color in the face area. However, it was not verified against a reference signal and the accuracy of the estimated heart rate was not quantitatively examined. In [51, 57, 58], the authors built resilience against the motion induced illumination changes by finding a particular direction in the 3-D space of the RGB channels that was the least affected. They exploited the fact that illumination changes due to the face motion and the heartbeat have different causes, and analyzed using light reflection characteristics and face-camera geometry. However, their use of the Viola–Jones face detector [59] and/or face-level affine transform did not precisely align faces at the pixel level, which could add noise to the extracted face color signals.

In this chapter, we aim to examine the best possible performance for fitness exercise videos when the registration error is minimized for the color-based heart-rate monitoring method. A block diagram of our proposed method is shown in Fig. 5.1. We minimize the registration error using pixel-level optical flow based motion compensation [60, 61] that is capable of generating almost "frozen" videos for best extracting the face color signals. We use the RGB weights proposed in [57] to resist unwanted illumination changes due to motion. We focus on the fitness scenarios that heart rate often wildly vary at different stages of fitness exercises, and present our results in widely adopted metrics [49, 55, 62] for comparison purpose.

The rest of the chapter is organized as follows. In Chapter 5.2, we propose our video-based heart-rate monitoring method specially designed for fitness exercises. In Chapter 5.3, we present the experimental results with comparisons if some modules

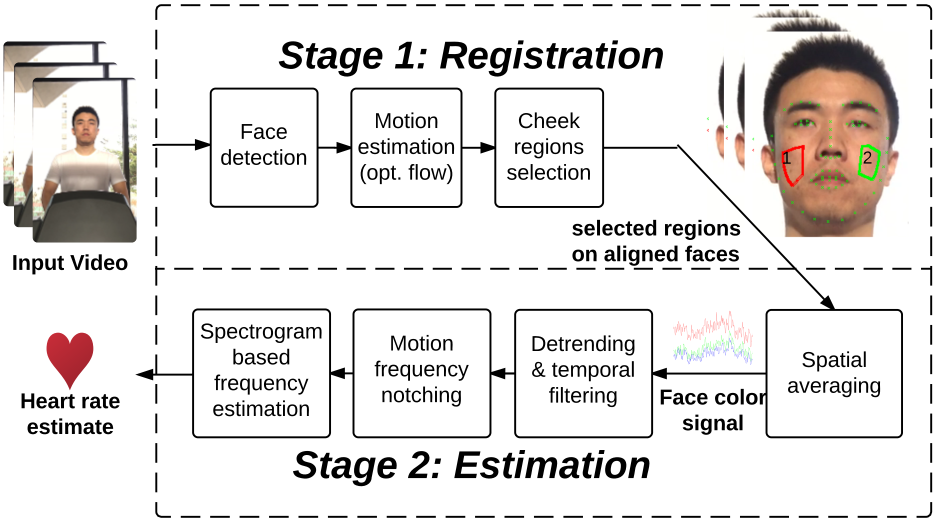

Stage 1: Registration
Input Video
Face detection
Motion estimation (opt. flow)
Cheek regions selection
1
2
selected regions on aligned faces
Spectrogram based frequency estimation
Motion frequency notching
Detrending & temporal filtering
Spatial averaging
Face color signal
Heart rate estimate
Stage 2: Estimation

Figure 5.1: Flowchart for the proposed heart rate monitoring method for fitness exercise videos.

were otherwise replaced or turned off. Finally, Chapter 5.4 concludes the chapter.

5.2 Proposed Method

Fitness exercise videos may contain large and periodic motions. Our proposed method focuses on a highly precise motion compensation scheme to allow generating a clean face color signal to facilitate the latter analysis steps, and uses the resulting motion cue as the guide to adaptively remove ambiguous frequency components that can be very close to the heart rate.

5.2.1 Precise Face Registration

A highly precise pixel-level motion compensation is a crucial step toward generating a clean face color signal. We use an optical flow algorithm to find correspondences of all points on the faces between two frames. Optical flow uses gradient information to iteratively refine the estimated motion vector field [60]. To avoid being trapped in local optima, we introduce a prealignment stage to bring the face images roughly aligned before conducting a fine-grain alignment using optical flow.

We use the Viola–Jones face detector [59] to obtain rough estimates of the location and size of the face. We clip and resize the face region of each frame to 180 pixels in height, effectively generating a prealigned video for the face region.

The prealignment significantly reduces the lengths of motion vectors, which in turn makes results of optical flow more reliable. In our problem, two face images are likely have a global color difference due to the heartbeat. In order to conduct a precise face alignment, instead of using the illumination consistency assumption that is widely used, we assume more generally that the intensity I of a point in two frames are related by an affine model, namely,

$$I(x + \Delta x, y + \Delta y, t + 1) = (1 - \epsilon)\, I(x, y, t) + b \tag{5.1}$$

where ϵ and b control the scaling and bias of the intensities between two frames. Both of them are usually small. Traditional techniques tackling the illumination consistency cases such as Taylor expansion and regularization can be similarly applied. Our mathematical analysis showed that omitting the illumination change due to the heartbeat, and applying a standard optical flow method leads to a bias term that is at the same order magnitude compared to the intrinsic error (in terms of standard deviation) of the optical flow system. We therefore use Liu's optical flow implementation [61] in our work.

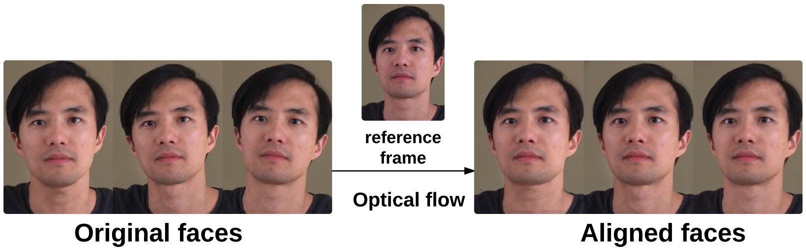


Figure 5.2: Face images from a same video segment before and after optical flow based motion compensation using the same reference face.

We divide each video into small temporal segments with one frame overlapping for successive segments. We use the frame in the middle of the segment as the reference for optical flow based motion compensation. This would ensure two frames being aligned do not have significant occlusion due to long separation in time. Fig. 5.2 shows a couple of face images from a same segment before and after optical flow based motion compensation using the same reference.

5.2.2 Segment Continuity and Cheek Regions Selection

With the precisely aligned face videos in short segments, we can estimate the face color for each frame by taking a spatial average over pixels of the cheek for R, G, and B channels, respectively. We call the three resulting 1-D time signals the *face color signals.*

When concatenating segments into color signals, the last point of the current segment and the first point of the next segment may have different intensities because they correspond to the same frame whose motion compensation were conducted with respect to two different references. To address this problem, the difference of the intensity between the two points is calculated and the resulting value is used to bias the signal of the next segment in order to maintain the continuity.

The face color signals contain color change due to the heartbeat, and illumination change due to face motions such as tilting. The green channel was used because it corresponds to the absorption peak of (oxy-) hemoglobin [44] that changes periodically as the heartbeat, and source separation methods such as the independent component analysis (ICA) were also used to separate the heartbeat component [46]. In [57], the authors proposed using the fixed linear weights $(-1, 2, -1)$ for R, G, B channels to best retain the heartbeat component while compensating the motion induced illumination change. In our experiments, we found that the fixed weights approach outperforms all other approaches, and we therefore adopt it in our proposed method.

To determine the cheek regions for conducting spatial averaging, we construct two conservative regions that do not contain facial structures and are most upfront in order to avoid strong motion-induced specular illumination changes. We use facial landmarks identified by the method proposed in [63] to facilitate the construction of the cheek regions. Each cheek region is constructed to be a polygon that has a safe margin to the facial structures protected by the landmarks. One example for such selected cheek regions and corresponding face landmarks is shown on the face in Fig. 5.1.

5.2.3 Detrending and Temporal Filtering

Illumination variation caused by passersby and/or the gradual change of sun light can cause the face color signal to drift, which is problematic for Fourier-based analysis. Such slowly-varying trend can be estimated and then subtracted from a raw face color signal, $\mathbf{x}_{\text{raw}} \in \mathbb{R}^L$, where L is the length of the signal. The trend is assumed to be a clean, unknown version of $\mathbf{x}_{\text{raw}}$ with a property that its accumulated convexity measured for every point on the signal is as small as possible, namely,

$$\hat{\mathbf{x}}_{\text{trend}} = \underset{\mathbf{x}}{\operatorname{argmin}} ||\mathbf{x}_{\text{raw}} - \mathbf{x}||^2 + \lambda||\mathbf{D}_2\mathbf{x}||^2 \tag{5.2}$$

where λ is a regularization parameter controlling the smoothness of the estimated trend, and $\mathbf{D}_2 \in \mathbb{R}^{L \times L}$ is a spare toeplitz second-order difference matrix. The closed-form solution is $\hat{\mathbf{x}}_{\text{trend}} = (\mathbf{I} + \lambda\mathbf{D}_2^T\mathbf{D}_2)^{-1}\mathbf{x}_{\text{raw}}$. Hence, the detrended signal is $\mathbf{x}_{\text{raw}} - \hat{\mathbf{x}}_{\text{trend}}$.

After detrending, we use a bandpass filter to reject the frequency components that are outside a normal range of human heart rate. The filter of choice is an IIR Butterworth with a passband from 40 to 240 bpm.

5.2.4 Motion Frequency Notching

In previous stages, we have designed our method to best remove the impact of face motions: optical flow was used to precisely align the faces, and a color weight vector that is least susceptible to motion was used to reduce impact of the periodic illumination change due to the face tilting. In this part, we further apply a notching operation to remove any remaining trace. We combine motion information from the face tracker and the optical flow to generate two time signals, one for the x-direction and the other for the y-direction. For each time bin on the spectrogram, we conduct two notch operations based on the dominating frequency estimated from the x- and y- motion components.

Spectrograms in the first column of Fig. 5.4 show that motion traces exist before notching, as highlighted by the arrows. We notice that the motion artifacts can be even stronger than the heart rate (HR) traces. Spectrograms in the second column of Fig. 5.4 show that the frequency notching method is effective and the HR traces dominate after notching.

5.2.5 Robust Frequency Estimation

We design a robust frequency estimator for noisy face color signals from fitness exercises. Instead of directly finding the peak (the mode) of the power spectrum for every time bin that may result in a discontinuous estimated heart-rate signal, we construct a two-step process to ensure the estimated signal is smooth.

We first find a single most probable strap from the spectrogram. We binarize each time bin of the spectrogram image per the 95th percentile of the power spectrum of that bin. We then dilate and erode the image in order to connect the broken strap. We find the largest connected region using such standard traverse algorithm as the breadth-first search and consider it as the most probable strap. A spectrogram and the results of successive steps are shown in Fig. 5.3.

We finally use a weighted frequency [64] within the the frequency range specified by the strap, $\mathcal{F}_i$, as the frequency estimate for ith time bin. Denoting the frequency estimate as $\hat{f}_{\mathrm{HR}}(i)$, we have

$$\hat{f}_{\mathrm{HR}}(i) = \sum_{f \in \mathcal{F}_i} w_{i,f} \cdot f \tag{5.3}$$

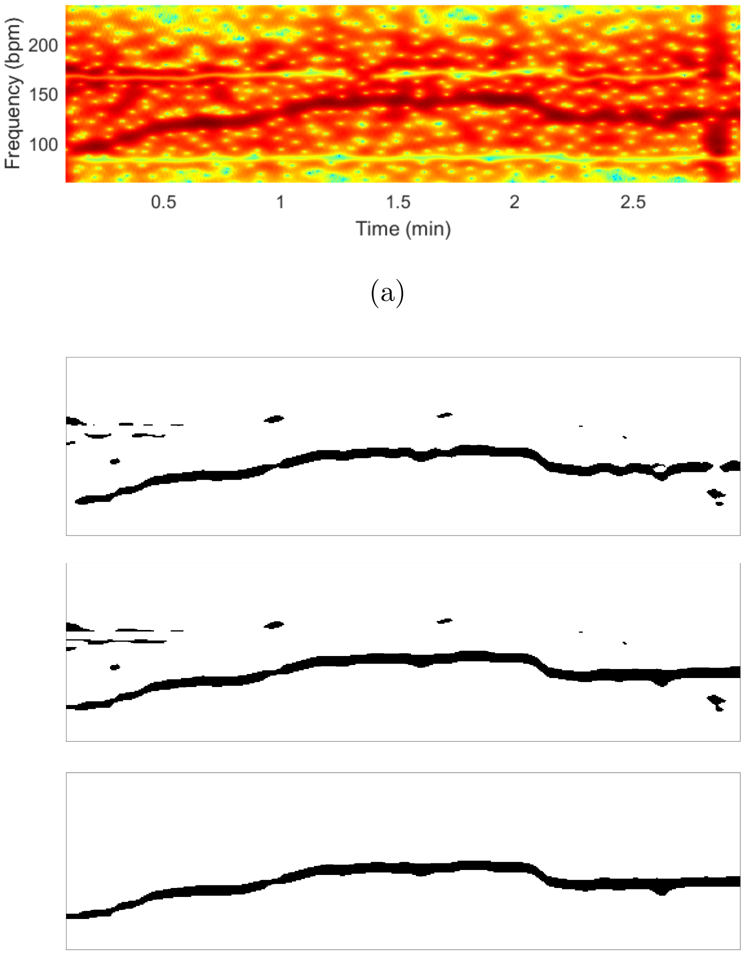
where $w_{i,f} = |S(i,f)| \,/\, \sum_{f \in \mathcal{F}_i} |S(i,f)|$, and $S(i,:)$ is the power spectrum at the ith bin.

5.3 Experimental Results

Our proposed method was evaluated on a self-collected fitness exercise dataset to demonstrate the efficacy on dealing with fitness motions, and results were presented in widely adopted metrics [49,55,62] in the field of heart rate monitoring from videos.

The fitness exercise dataset has 9 videos in which 6 contain human motions on an elliptical machine and the other 3 contain motions on a treadmill. Each video is about 3 minutes long in order to cover various stages of a fitness exercise. Each video was captured in front of the face by a commodity mobile camera (iPhone 6s) affixed on a tripod or held by the hands of a person other than the test subject. The gym was well-lit with several over-the-top florescent lights and with diffuse daylight passing into the gym through glass walls. The heart rate of the test subject was simultaneously monitored by an electrocardiogram (ECG)-based chest belt (Polar H7) for reference.

Each video was divided into segments of 1.5 secs in order to guarantee small scene changes within each segment for optical flow's best performance. The regu-



(a)

(b)

Figure 5.3: (a) A spectrogram with weakly connected frequency strap. (b) Results after the following operations (from top to bottom): binarization using 95th percentile, dilation and erosion, and small regions removal.

larization parameter for the detrending on the face color signal was set to $\lambda = 20$ for 30Hz videos used in this experiment. The window length for spectrogram was set to 10 secs with 98% overlap.

Representative results from two videos are shown in Fig. 5.4. Row 1 and row 2 show the spectrograms for the detrended and filtered face color signal before and

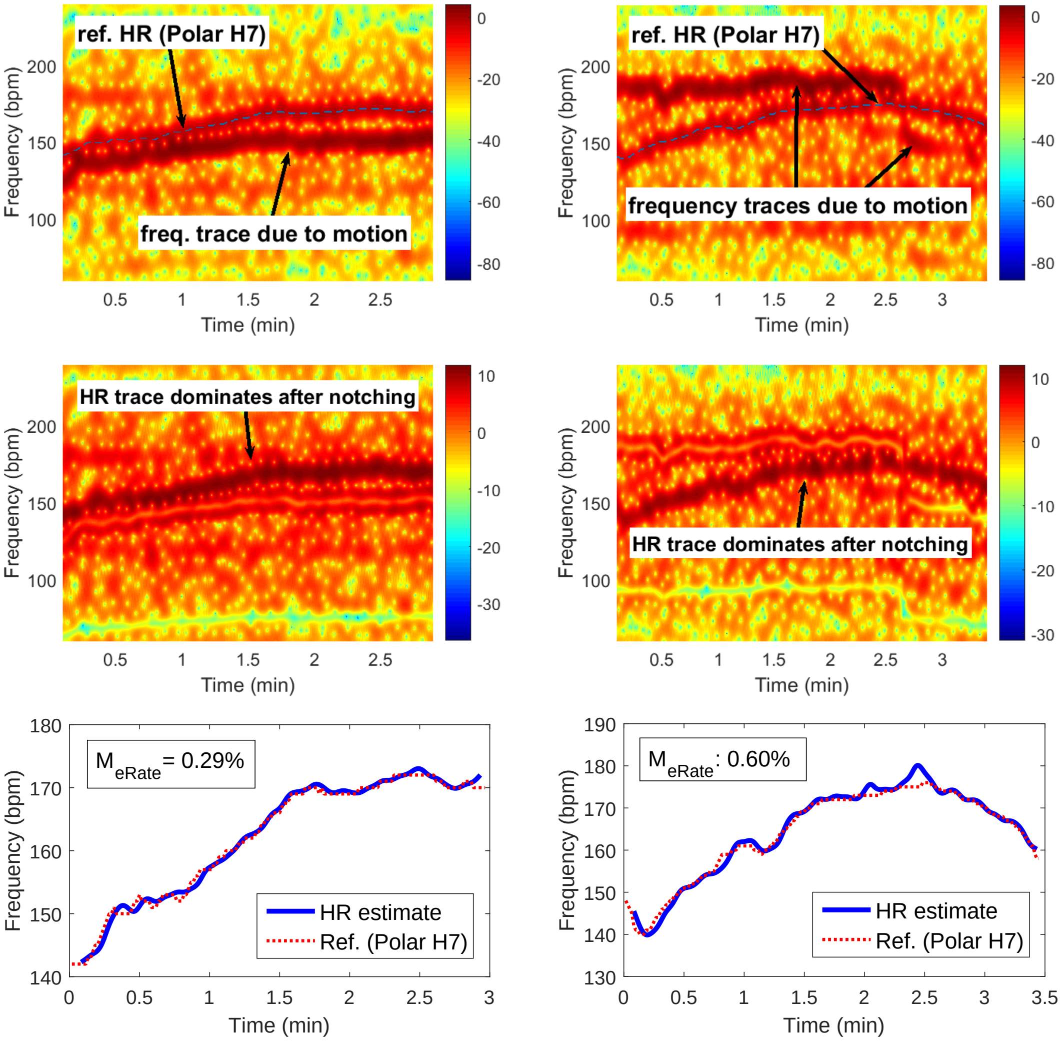


Figure 5.4: Contrast of spectrograms before (row 1) and after (row 2) notching the frequencies of fitness motions. Row 3: Heart rate estimates when compared to the ECG-based reference measurement using Polar H7 chest belt. Left column: from elliptical machine, URL: http://goo.gl/e0WCnc; Right column: from treadmill, URL: http://goo.gl/8GLoLB.

after motion information guided notching. Row 3 shows the HR estimates obtained using the robust frequency estimation algorithm. We plotted the HR estimates with the reference HR, and found that the estimates are almost unbiased and are

Table 5.1: Performance (in terms of mean and standard deviation in parentheses) of proposed method and cases when some modules were otherwise replaced or turned off.

Module combinations	RMSE in bpm	M_{eRate}
tracker + JBSS (no op)	7.6 (5.7)	3.60% (2.87%)
tracker + fixed (no op)	5.6 (3.4)	2.61% (1.45%)
tracker + op + JBSS	1.3 (0.7)	0.65% (0.30%)
tracker + op + fixed (proposed)	**1.1 (0.6)**	**0.58% (0.33%)**

fluctuating around the reference. The relative error (M_{eRate}) are as low as 0.29% and 0.60% for the two videos, respectively. We included two demos, each of which contains a raw video, a motion compensated video, and a synchronized HR estimate and a reference HR.

We summarize the mean and standard deviation of the error measures for all of our videos and the results are listed in Table 5.1. The averaged error for the proposed method is 1.1 bpm in root mean-squared error (RMSE) and 0.58% in relative error. The performance is slightly reduced if a more complicated joint blind-source separation (JBSS) approach is used to search for the weights for R, G, B channels, instead of using the more theoretically grounded fixed weights $(-1, 2, -1)$.

We conducted additional experiments to check the impact when the optical flow algorithm was disabled. In this case, the face images were roughly aligned using the face tracker. The reported errors in Table 5.1 show significant performance reduction from 1.1 bpm to 5.6 bpm or from 0.58% to 2.61%. The estimation error has increased about four times, which shows that a precise alignment is a crucial step for the video-based heart-rate monitoring method for fitness scenarios.

5.4 Chapter Summary

In this chapter, we proposed a heart rate monitoring method for fitness exercise videos. We focused on building a highly precise motion compensation scheme with the help of the optical flow, and used motion information as a cue to adaptively remove ambiguous frequency components for improving the heart rates estimates. Experimental results show that our proposed method can give precise estimates at an average error of 1.1 bpm in RMSE or 0.58% in relative error.

6

Conclusions and Future Perspectives

In this dissertation, we have explored micro signal extraction with a focus on the visual aspect using mobile devices. We have tackled three micro signal extraction problems, whose challenge lies either in the relative magnitude or in the topological scale, by synergistically applying and adapting signal processing theories and techniques.

For micro signal extraction in forensics applications, we have investigated intrinsic microscopic feature of the paper surface for authentication purpose. We have shown that it is possible to use the cameras and built-in flashlights of mobile devices to estimate the normal vector field of paper surfaces. Perturbation analysis shows that the proposed method is robust to inaccurate estimates of camera locations, and using 6 to 8 images can achieve a matching accuracy of 10^{-4} in EER under a lab-controlled ambient light environment. This finding can relax the restricted imaging setup in prior art, and enable paper authentication under a more casual, ubiquitous setting with a mobile imaging device. The proposed technique may facilitate duplicate detection of important and/or valuable documents such as IDs, and facilitate counterfeit mitigation of merchandise via detection of duplicated labels and packages.

Our results from fitted probability density functions show that the uniqueness of the surfaces is satisfactory in term of ROCs measure for practical identification applications. It is yet of theoretical interest to understand how much inherent randomness there exists in the normal vector field. The entropy widely used in the fingerprint community can be one candidate measure. It can be beneficial to model the normal vector field using a generative random field to gain further understanding.

The proposed identification method is based on flat surfaces, and it can be of practical interest to extend the method to non-flat, but parametrically shaped surfaces. Investigation in this scenario will not only allow the proposed method work on flat surfaces such as papers and boxes, but also on wine bottles and other parametrically shaped products. The current assumption of perspective distortion can be generalized to compensate higher order geometric distortions.

The proposed authentication method is made possible by using a flashlight for multiple images. The flashlight can be visually unpleasant for some users, and may not be permitted in certain scenarios. If one can estimate the ambient light, and use the estimated ambient light for normal vector field estimation, the user-friendliness of the authentication method can be improved.

We have also explored the use of an invisible power signature, the ENF signal, that may be embedded in images taken by CMOS cameras for geo-tagging purposes. Specifically, we addressed the research question of whether the basic attributes of an embedded ENF signal can be correctly identified. We proposed two algorithms, the ICA-based method and the entropy minimization method. Experimental results show that both methods enable making highly accurate decisions when the ENF traces are strong, whereas the entropy minimization method outperforms the ICA-based method as ENF traces become weaker. In this proof-of-concept work, we have demonstrated a unique computer vision capability of extracting invisible traces to help narrow the capturing geographic region of an image.

The next step would be to examine and improve the performance on a larger scale investigation with more cameras and images, and explore additional location information that can be inferred from the ENF traces in an image. A more challenging problem is to estimate the exact value of the ENF frequency, instead of merely deciding between 50Hz and 60Hz. A precise estimator can benefit the time-location search with the presence of the reference ENF signals from various power grids.

For micro signal extraction in healthcare monitoring applications, we have proposed a heart rate monitoring method using fitness exercise videos. We focused on building a highly precise motion compensation scheme with the help of the optical flow, and used motion information as a cue to adaptively remove ambiguous frequency components for improving the heart rate estimates. Experimental results show that our proposed method can give precise estimates with an average error of 1.1 bpm in RMSE or 0.58% in relative error.

It would be interesting to explore the possibility of recovering some features of the ECG signal from the sinusoid-like heart-rate signal extracted from the face. Both ECG and the heart-rate signal from the face result from the heartbeat, with ECG containing much more diagnostic information. It is beneficial to collect simultaneous measurement of the ECG signal and the face color change signal in order to reveal a possible relationship between the two.

Bibliography

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

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

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

[4] R. Garg, A. L. Varna, A. Hajj-Ahmad, and M. Wu, "'Seeing' ENF: Power-signature-based timestamp for digital multimedia via optical sensing and signal processing," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 9, pp. 1417–1432, September 2013.

[5] M. C. Stamm, M. Wu, and K. J. R. Liu, "Information forensics: An overview of the first decade," *IEEE Access*, vol. 1, pp. 167–200, May 2013.

[6] R. Szeliski, *Computer Vision: Algorithms and Applications.* Springer, 2010.

[7] G. Kortuem, F. Kawsar, V. Sundramoorthy, and D. Fitton, "Smart objects as building blocks for the internet of things," *IEEE Internet Computing*, vol. 14, no. 1, pp. 44–51, Jan 2010.

[8] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Computer Networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[9] D. Miorandi, S. Sicari, F. D. Pellegrini, and I. Chlamtac, "Internet of things: Vision, applications and research challenges," *Ad Hoc Networks*, vol. 10, no. 7, pp. 1497–1516, 2012.

[10] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of things (IoT): A vision, architectural elements, and future directions," *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645–1660, 2013.

[11] Product Overview on BubbleTag™, Ramdot™, FiberTag™, *Prooftag SAS*, Retrieved Jan. 2015. http://www.prooftag.net/

[12] Kinde Anti-Counterfeiting Labels, *Guangdong Zhengdi (Kinde) Network Technology Co., Ltd.*, Retrieved Jan. 2015. http://www.kd315.com/

[13] C.-W. Wong and M. Wu, "A study on PUF characteristics for counterfeiting detection," in *Proc. IEEE International Conference on Image Processing (ICIP)*, Quebec City, Canada, Sep. 2015, pp. 1643–1647.

[14] W. Clarkson, T. Weyrich, A. Finkelstein, N. Heninger, J. Halderman, and E. Felten, "Fingerprinting blank paper using commodity scanners," in *Proc. IEEE Symposium on Security and Privacy*, Berkeley, CA, May 2009, pp. 301–314.

[15] S. Voloshynovskiy, M. Diephuis, F. Beekhof, O. Koval, and B. Keel, "Towards reproducible results in authentication based on physical non-cloneable functions: The forensic authentication microstructure optical set (FAMOS)," in *Proc. IEEE International Workshop on Information Forensics and Security (WIFS)*, Tenerife, Spain, Dec. 2012, pp. 43–48.

[16] M. Diephuis and S. Voloshynovskiy, "Physical object identification based on FAMOS microstructure fingerprinting: Comparison of templates versus invariant features," in *Proc. International Symposium on Image and Signal Processing and Analysis (ISPA)*, Trieste, Italy, Sep. 2013, pp. 119–123.

[17] M. Diephuis, S. Voloshynovskiy, T. Holotyak, N. Stendardo, and B. Keel, "A framework for fast and secure packaging identification on mobile phones," in *Proc. SPIE, Media Watermarking, Security, and Forensics*, San Francisco, CA, Feb. 2014, p. 90280T.

[18] S. A. Shafer, "Using color to separate reflection components," *Color Research & Application*, vol. 19, no. 4, pp. 210–218, 1986.

[19] C.-W. Wong and M. Wu, "Counterfeit detection using paper PUF and mobile cameras," in *Proc. IEEE International Workshop on Information Forensics and Security (WIFS)*, Rome, Italy, Nov. 2015.

[20] "High resolution surface topography FRT MicroProf chromatic aberration sensor," Aug. 2012, a product sheet by Innventia AB.

[21] S. K. Nayar, K. Ikeuchi, and T. Kanade, "Surface reflection: Physical and geometrical perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 7, pp. 611–634, Jul. 1991.

[22] B. C. Levy, *Principles of Signal Detection and Parameter Estimation.* Springer, 2008.

[23] N. Poh and S. Bengio, "How do correlation and variance of base-experts affect fusion in biometric authentication tasks?" *IEEE Transactions on Signal Processing*, vol. 53, no. 11, pp. 4384–4396, Nov. 2005.

[24] P. J. Huber and E. M. Ronchetti, *Robust Statistics.* Wiley, 2009.

[25] C. E. McCulloch and S. R. Searle, *Generalized, Linear, and Mixed Models*. Wiley, 2001.

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

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

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

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

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

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

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

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

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

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

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

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

[38] A. Hajj-Ahmad, S. Baudry, B. Chupeau, and G. Doërr, "Flicker forensics for pirate device identification," in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, ser. IH&MMSec. Portland, Oregon, USA: ACM, 2015, pp. 75–84. http://doi.acm.org/10.1145/2756601.2756612

[39] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis.* John Wiley and Sons, 2002.

[40] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* John Wiley and Sons, 2006.

[41] T. Kim, T. Eltoft, and T.-W. Lee, *Independent Vector Analysis: An Extension of ICA to Multivariate Components.* Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 165–172.

[42] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation." *Opt. Express*, vol. 18, no. 10, pp. 10 762–10 774, May 2010.

[43] J. Cheng, X. Chen, L. Xu, and Z. J. Wang, "Illumination variation-resistant video-based heart rate measurement using joint blind source separation and ensemble empirical mode decomposition," *IEEE Journal of Biomedical and Health Informatics*, 2016.

[44] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light," *Optices Express*, vol. 16, no. 26, pp. 21 434–45, Dec. 2008.

[45] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation." *Optics express*, vol. 18, no. 10, pp. 10 762–74, May 2010.

[46] ——, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 7–11, Jan. 2011.

[47] M. Lewandowska, J. Rumiski, T. Kocejko, and J. Nowak, "Measuring pulse rate with a webcam – a non-contact method for evaluating cardiac activity," in *Federated Conference on Computer Science and Information Systems (FedCSIS)*, Szczecin, Poland, Sep. 2011, pp. 405–410.

[48] S. Kwon, H. Kim, and K. S. Park, "Validation of heart rate extraction using video imaging on a built-in camera system of a smartphone," in *IEEE EMBS Annual International Conference*, San Diego, CA, Aug. 2012, pp. 2174–2177.

[49] X. Li, J. Chen, G. Zhao, and M. Pietikainen, "Remote heart rate measurement from face videos under realistic situations," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, Jun. 2014, pp. 4264–4271.

[50] Y. Cui, C.-H. Fu, H. Hong, Y. Zhang, and F. Shu, "Non-contact time varying heart rate monitoring in exercise by video camera," in *International Conference on Wireless Communications & Signal Processing (WCSP)*, Nanjing, China, Oct. 2015.

[51] L. Feng, L. M. Po, X. Xu, Y. Li, and R. Ma, "Motion-resistant remote imaging photoplethysmography based on the optical properties of skin," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 5, pp. 879–891, May 2015.

[52] S. Yu, X. You, X. Jiang, K. Zhao, Y. Mou, W. Ou, Y. Tang, and C. L. P. Chen, "Human heart rate estimation using ordinary cameras under natural movement," in *IEEE International Conference on Systems, Man, and Cybernetics*, Hong Kong, Oct. 2015, pp. 1041–1046.

[53] W. Wang, S. Stuijk, and G. de Haan, "Exploiting spatial redundancy of image sensor for motion robust rPPG," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 2, pp. 415–425, Feb. 2015.

[54] S. Fernando, W. Wang, I. Kirenko, G. de Haan, S. Bambang Oetomo, H. Corporaal, and J. van Dalfsen, "Feasibility of contactless pulse rate monitoring of neonates using Google Glass," in *EAI International Conference on Wireless Mobile Communication and Healthcare (MOBIHEALTH)*, ICST, Brussels, Belgium, Belgium, 2015, pp. 198–201.

[55] S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J. F. Cohn, and N. Sebe, "Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, Jun. 2016, pp. 2396–2404.

[56] Y. Sun, S. Hu, V. Azorin-Peris, S. Greenwald, J. Chambers, and Y. Zhu, "Motion-compensated noncontact imaging photoplethysmography to monitor cardiorespiratory status during exercise," *Journal of Biomedical Optics*, vol. 16, no. 7, pp. 077 010:1–9, Jul. 2011.

[57] G. de Haan and V. Jeanne, "Robust pulse rate from chrominance-based rPPG," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 10, pp. 2878–2886, Oct. 2013.

[58] G. de Haan and A. van Leest, "Improved motion robustness of remote-PPG by using the blood volume pulse signature," *Physiological Measurement*, vol. 35, no. 9, p. 1913, Aug. 2014.

[59] P. Viola and M. J. Jones, “Robust real-time face detection,” *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, May 2004.

[60] B. D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” in *International Joint Conference on Artificial Intelligence*, vol. 2, Vancouver, Canada, Aug. 1981, pp. 674–679.

[61] C. Liu, “Beyond pixels: exploring new representations and applications for motion analysis,” Ph.D. dissertation, Massachusetts Institute of Technology, 2009.

[62] M. A. Haque, R. Irani, K. Nasrollahi, and T. B. Moeslund, “Heartbeat rate measurement from facial video,” *IEEE Intelligent Systems*, vol. 31, no. 3, pp. 40–48, May 2016.

[63] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas, “Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model,” in *IEEE International Conference on Computer Vision (ICCV)*, Sydney, Australia, Dec. 2013, pp. 1944–1951.

[64] R. Garg, A. L. Varna, A. Hajj-Ahmad, and M. Wu, “‘Seeing’ ENF: Power-signature-based timestamp for digital multimedia via optical sensing and signal processing,” *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 9, pp. 1417–1432, Sep. 2013.