ABSTRACT

Title of dissertation: TEMPORAL AND SPATIAL ALIGNMENT OF MULTIMEDIA SIGNALS

Hui Su, Doctor of Philosophy, 2014

Dissertation directed by: Professor Min Wu
Department of Electrical and Computer Engineering

With the increasing availability of cameras and other mobile devices, digital images and videos are becoming ubiquitous. Research efforts have been made to develop technologies that utilize multiple pieces of multimedia information simultaneously. This dissertation focuses on the temporal and spatial alignment of multimedia signals, which is a fundamental problem that needs to be solved to enable such applications dealing with multiple pieces of multimedia data.

The first part of the dissertation addresses the synchronization of multimedia signals. We propose a new modality for audio and video synchronization based on the electric network frequency (ENF) signal naturally embedded in multimedia recordings. Synchronization of audio and video is achieved by aligning the ENF signals. The proposed method offers a significant departure to tackling the audio/video synchronization problem from existing work, and a strong potential to address previously untractable scenarios.

Estimation of the ENF signal from video is a challenging task. In order to address the problem of insufficient sampling rate of video, we propose to exploit the

rolling shutter mechanism commonly adopted in CMOS camera sensors. Several techniques are designed to alleviate the distortions of motions and brightness changes in videos for ENF estimation.

We also address several challenges that are unique to the synchronization of digitized analog audio recordings. Speed offset often occurs in digitized analog audio recordings due to the inconsistency in the tape's rolling speed. We show that the ENF signal captured by the original analog audio recording can be retained in the digitized version. The ENF signal is considered approximately as a single-tone signal and used as a reference to detect and correct speed offsets automatically.

A complete multimedia application system often needs to jointly consider both temporal synchronization and spatial alignment. The last part of the dissertation examines the quality assessment of local image features for efficient and robust spatial alignment. We propose a scheme to evaluate the quality of SIFT features in terms of their robustness and discriminability. A quality score is assigned to every SIFT feature based on its contrast value, scale and descriptor, using a quality metric kernel that is obtained in a one-time training phase. Feature selection is performed by retaining features with high quality scores. The proposed approach is also applicable to other local image features, such as the Speeded Up Robust Features (SURF).

TEMPORAL AND SPATIAL ALIGNMENT OF MULTIMEDIA SIGNALS

by

Hui Su

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
2014

Advisory Committee:
Professor Min Wu, Chair/Advisor
Professor Carol Espy-Wilson
Professor Kari Kraus
Professor K. J. Ray Liu
Professor Gang Qu

To my family

Acknowledgments

First and foremost, I would like to express my sincere and greatest gratitude to my advisor, Prof. Min Wu, for her continuous support and guidance over the past five years of my graduate study. With her vision, creativeness, and enthusiasm, she has taught me how to identify important research issues, and how to solve problems with critical thinking and unremitting efforts. Her encouragement and constructive suggestions have led me through difficult times. Her endless pursuit of excellence inspired me to strive for achieving my potential in my study, research, and professional careers. The valuable lessons I learned from her not only fueled my graduate study and research, but will also have a lasting influence in my future professional and personal development.

I would like to thank Prof. K.J. Ray Liu for his enlightening courses and his constructive suggestions on my research. I am grateful to Prof. Kari Kraus for her expertise and insightful comments in our collaborative work. I want to thank Prof. Gang Qu and Prof. Carol Espy-Wilson for serving on my dissertation committee and for their valuable comments. I also want to thank Prof. Douglas Oard for the enlightening discussions that inspired our collaborative work.

It has been my pleasure and privilege to work with my wonderful colleagues in Maryland. I thank Dr. Avinash Varna, Dr. Wenjun Lu, Dr. Wei-Hong Chuang, and Dr. Ravi Garg for providing help and guidance on my research. I thank my officemates Chau-Wai Wong, Adi Hajj-Ahmad, and Abbas Kazemipour with whom I enjoyed our discussions on various research and life topics.

Last but not least, I thank my parents who always give me unconditional love and support. My gratitude and love to them are beyond words. I dedicate this dissertation to them.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Recent years have witnessed rapid growth in the amount of digital multimedia data. With the advancement in device design and manufacturing technologies, mobile phones and portable cameras with improved imaging quality and large memory storage capacity are becoming ubiquitous. These developments encourage people to take photos and audio/video recordings to capture interesting moments and important events. Photos created by different people can often include the same objects such as landmarks. Similarly, events like concerts or speeches may be recorded in multiple audio and video clips. Numerous applications arise for which the common content in images and audio/video recordings is exploited.

When an event is recorded simultaneously by multiple independent cameras, and possibly from a variety of angles, combining the information in these videos may provide a better presentation and novel experience of the event than any recording alone. For example, using 3D reconstruction techniques, a dynamic scene may be

reconstructed from multiple video streams that allows people to choose from different viewing angles of a scene. Several videos of various perspective can be "stitched" together to achieve wider field of view via video panorama [6]. A video sequence of high space-time resolution can be obtained by combining information from multiple low-resolution video sequences of the same dynamic scene [53].

Similarly, multiple photos containing the same object or scene can be exploited to create visually appealing images with panoramic views and super-fine resolutions. The existence of common objects in multiple images have also motivated research in content-based image retrieval (CBIR). Given a query image, the goal is to retrieve from a large database the images containing the same object or scene as in the query image. CBIR can be useful in scenarios including landmark recognition, direction aid for tourists, and CD/book automatic annotation.

1.1 Temporal and Spatial Alignment of Multimedia data

Temporal and spatial alignment are two essential problems to solve for many applications involving multiple pieces of audio-visual data. Temporal alignment, or synchronization, is the process of finding time correspondence between multiple audio/video streams. Most existing approaches to multimedia signal synchronization extract and match audio/visual features extracted from the audio/video recordings. These methods may not perform well in certain situations. For example, it is difficult to synchronize video sequences using visual features when the videos do not share sufficient common areas of the scene or when the viewing angles differ sig-

nificantly; similar limitations apply to alignment of audio recordings that have few common acoustic or speech components.

In addition to synchronization, a complete multimedia application system must also address spatial alignment, which is the process of transforming different images of the same scene or object into one coordinate system. Local image feature is one of the most successful solution for image alignment. In the local feature framework, interest points are first selected as distinctive and robust points in the image by a key point detector. Next, a robust feature descriptor is generated using the information within the neighborhood of the interest point. The interest point and feature descriptor are designed in a way so that the same object can produce interest points at the same image structure with similar descriptors regardless of the changes of viewing angles and lighting conditions. The point correspondence between images of the same scene can be obtained by matching the feature descriptors.

Most of the solutions for CBIR are also built on matching local image features. To cope with large-scale databases, a visual dictionary-based Bog-of-Words (BoW) approach has been proposed by Sivic and Zisserman [55]. They propose to quantize image features into a set of visual words as a codebook by using k-means clustering on many training features. A given feature can be mapped to its nearest visual word among the codebook. The images are then represented by the frequency vector of the visual word occurrences. The similarity between two images is usually measured using the L_1 or L_2 distance between their visual word frequency vectors. During a query, the similarity score can be computed efficiently by using an inverted file system associated to the database.

1.2 Main Contributions

We explore two main research problems related to the temporal and spatial alignment of multimedia data in this dissertation. The first problem asks how to synchronize audio and video recordings in adverse conditions for conventional methods. We propose a new synchronization modality that exploits the electric network frequency (ENF) signal naturally embedded in multimedia recordings. In contrast to conventional methods, our proposed approach does not rely on the perceptual audio and visual information of the recordings. It offers a strong potential to address difficult scenarios that are otherwise intractable. For video synchronization, it imposes no major constraints on the viewing angles and camera calibrations, as other methods do. The second problem we explored is how to improve the robustness and efficiency of feature-based image matching. Image feature matching remains critical for image alignment and retrieval. The improvement of image resolution and the growth of image database in scale may lead to the explosion of the number of local image features. To expedite feature matching without greatly affecting accuracy, it is desirable to select a subset of the most reliable and informative features to represent each image. We propose a method to evaluate feature quality in terms of robustness and discriminability. Every feature is assigned a quality score used for feature selection. The topics explored in this dissertation are listed below.

1.2.1 ENF as a New Modality for Synchronization

ENF is the supply frequency of the alternating current in a power grid. The nominal value of the ENF is 60Hz in the Americas, Taiwan, Saudi Arabia, and Philippines, and is 50Hz in other regions except Japan, which adopts both frequencies. The instantaneous value of the ENF constantly deviates slightly and randomly from its nominal value, as a result of the dynamic interaction between power generation and consumption and the control mechanism of the power grid. The instantaneous values of the ENF over time is referred to as the ENF signal. The variation patterns in the ENF signal are shown to be consistent within the same power grid, even for distant locations. Multimedia recordings created using electric devices plugged into the power mains or located near electric activities often pick up ENF signals in audio due to electromagnetic interference and/or acoustic vibrations [27]; and in video due to imperceptible flickering in indoor lighting [23].

Viewed as a continuous random process over time, the ENF signal embedded in audio and video signals can be used as a timing fingerprint that is unique at any specific time instance. In this dissertation, we propose to match the ENF signals extracted from audio/video recordings to achieve temporal alignment. After extracting ENF signals from the recordings to be synchronized, the normalized cross correlation coefficients are calculated with different lags between the ENF signals, and the lag corresponding to the maximum correlation coefficient is identified as the temporal shift between the recordings. The experimental results show that the synchronization error of the ENF-based method can be under 0.1 second.

1.2.2 Extraction of ENF from Videos Created with Rolling Shutter

ENF signals may be extracted from the soundtracks of the video recordings, as well as the image sequences if the video captures the subtle flickering of lights. Extracting the weak ENF signal from image sequences poses a challenging task which has not been well addressed. The temporal sampling rate of visual recordings is generally too low for estimating the ENF signal that may appear at harmonics of 50 or 60 Hz. The ENF traces in video signals are relatively weak, and they may easily be distorted by object and camera motions.

We exploit the rolling shutter that is commonly adopted for CMOS sensors to facilitate ENF estimation from video. Unlike global shutters often employed in charge-coupled device (CCD) sensors that record the entire frame from a snapshot of a single point in time, a camera with a rolling shutter scans the vertical or horizontal lines of each frame in a sequential manner. As a result, different lines in the same frame are exposed at slightly different times. By treating each line of the frame as a sample point, the sampling rate can be much higher than the frame rate, which helps solve the aliasing problem.

Existing work on ENF estimation from image sequence of video is limited to videos of static scenes. It is more challenging to extract ENF signals from video recordings with object motions, brightness changes, and camera motions. To address object motions, if the scene in the video contains a static background, we can use these static regions to estimate the ENF signal. Many cameras are equipped with a brightness control mechanism that adjusts the camera sensor's aperture and/or

sensitivity to light in response to the illumination conditions, so the overall brightness of the acquired image remains visually pleasing. Such brightness change affects the estimation of ENF signals. From our study, we find that the brightness changes may be well modeled by a linear transform. A linear transform is estimated to compensate for the brightness changes effectively. For videos with camera motions, optical flow-based method is used to estimate and compensate for the pixel shift among video frames.

1.2.3 Synchronization and Restoration of Old Audio Recordings

As ENF signal is embedded into multimedia recordings at the time of recording, several interesting challenges arise for ENF analysis on recaptured audio recordings. ENF signals in recaptured audio recordings may contain two components: one is inherited from the original recording, referred to as the *original ENF* signal; and the other becomes embedded during recapturing process, referred to as the *recapturing ENF* signal. When the original ENF and recapturing ENF overlap, conventional ENF estimation methods may fail. To solve this problem, a decorrelation based algorithm is designed to estimate effectively both ENF signals from recaptured audio signals in a sequential order.

In addition to the possible superposition of multiple ENF traces, the "drifting effect" present another distortion of the ENF signals unique to digitized analog tape recordings. Due to mechanical imperfection of analog recorders and tapes, the rolling speed of these tapes often varies over time during recording and replay. The

inconsistency between the rolling speeds of original recording and playback during digitization induce speed errors in the digitized audio. Techniques are designed to compensate for such drifting effect for accurate ENF analysis.

We also show that ENF signals suffering drifting effect can be exploited to correct tape speed errors. Using the ENF signal as a reference, the speed of audio can be restored by temporally stretching or compressing the audio signal via rate conversion. We have considered two schemes of rate conversion based on resampling and interpolation, respectively, and their performances are experimentally compared. The effectiveness of the proposed scheme is shown by a demonstration of speed restoration with recordings of the NASA Apollo 11 Mission.

1.2.4 Quality Assessment of Image Features For Efficient and Robust Visual Matching

Feature selection is required to handle image matching efficiently over large-scale data sets. We propose a quality metric to evaluate the usefulness of individual image features. Features with low quality scores may be discarded to reduce the number of features and to improve matching efficiency. We demonstrate the proposed approach with the SIFT feature, which is widely used in computer vision and image processing applications. The methodology may be generalized to other features as well.

Our feature quality metric uses a learning-based method. A feature space is established that consists of the scale, contract value, and descriptors of SIFT features.

The whole feature space is quantized into a certain number of bins. To improve performance, we adopt a soft quantization scheme. We collect a set of representative sample images covering various scene types and categories as training images, and extract SIFT features from these training images. The SIFT features are mapped to their nearest bins in the feature space. We conduct synthetic transformations of interest to the training images, match SIFT features between every original and transformed image pair, and record the matching results for the features in each feature space bin. The ratio of the number of correct matches over the total number of features in each bin is calculated and used as that bin's quality score. A quality metric kernel is thus obtained. Given a SIFT feature, we can compute its quality score using the quality metric kernel. A few nearest feature space bins are first identified, then the quality score of the feature is computed as the weighted average of the quality scores of its nearest bins.

The proposed approach is tested on three benchmark datasets for large scale content-based image retrieval. Feature selection according to the proposed quality score is shown to perform better than empirical methods, such as selecting features with largest scales and highest contrast values.

1.3 Dissertation Organization

The rest of the dissertation is organized as follows. In Chapter 2, we discuss how the ENF signal can be estimated from audio signals and used for synchronization of audio and video clips. Experiments on various audio and video recordings

are conducted to demonstrate the effectiveness of the proposed approach. The synchronization accuracy is also tested and analyzed.

In Chapter 3, we study the problem of ENF extraction from visual recordings, which is more challenging than audio recordings. The rolling shutter mechanism is first modeled and analyzed using multi-rate signal processing theory. To move beyond static videos, we propose several techniques to address the object motions, camera motions and brightness changes in video. We then conduct experiments to demonstrate that matching the ENF signals estimated from the image sequences of videos can effectively synchronize multiple video streams.

In Chapter 4, we investigate the ENF analysis on recaptured audio recordings. A decorrelation based method is introduced to estimate multiple ENF signals that may superimpose in a recaptured audio recording. The drifting effect of the ENF signal in digitized audio recording with speed errors is examined. We show that the ENF signal may serve as a reference signal to correct speed errors. Several historical recordings of the NASA Apollo 11 Mission are successfully restored with this method.

In Chapter 5, we investigate the SIFT feature selection for efficient and robust visual matching. A quality metric considering the scale, contrast value, and the descriptor of SIFT features is obtained through a learning based method. Feature selection using the proposed quality metric is shown to be effective on several benchmark data sets for large scale image retrieval. We also show that the methodology applies to other image local features such as SURF.

Finally, in Chapter 6, we conclude this dissertation and outline research issues

that can be explored in the future.

Chapter 2

ENF Signal as a New Modality for Video Synchronization

2.1 Chapter Introduction

The analysis of electric network frequency (ENF) signals has emerged in recent years as an important technique for digital multimedia forensics [22, 25, 27, 30–32, 50, 51, 58]. ENF is the supply frequency of the alternating current in a power grid. The nominal value of the ENF is 60Hz in the Americas, Taiwan, Saudi Arabia, and Philippines, and is 50Hz in other regions except Japan, which adopts both frequencies. The ENF has three interesting properties. First, the instantaneous value of the ENF constantly deviates slightly and randomly from its nominal value, as a result of the dynamic interaction between power generation and consumption and the control mechanism of the power grid. The instantaneous values of the ENF over time is referred to as the ENF signal. In second property of ENF, the

variation patterns in its instantaneous values are consistent within the same power grid, even for distant locations. For example, there are only three major power grids in the United States, the Eastern Grid, the Western Grid, and the Texas or ERCOT (Electric Reliability Council of Texas) Grid. So, there are mainly three different instantaneous variation patterns of ENF in the United States. Last but not least, multimedia recordings created using electric devices plugged into the power mains or located near electric activities often contain ENF signals in audio due to electromagnetic interference and/or acoustic vibrations [27]; and in video due to imperceptible flickering in indoor lighting [23]. The ENF signal extracted from audio or video recordings has been shown to exhibit a high correlation with the ENF extracted from the power mains measurements at the corresponding time.

Several forensic applications have been proposed based on the analysis of the ENF signal. For example, ENF signals have been successfully used as a natural time stamp to authenticate audio recordings [27, 32, 51]. By examining the phase continuity of the ENF signal, one can detect the region of tampering [50]. Some recent work shows that the ENF signal can also reveal information about the locations and regions in which certain recordings are made [22, 30, 31].

In this work, we explore the potential of the ENF signal from a new perspective and use it for synchronization of multimedia signals, i.e. the temporally alignment of audio and video recordings. Synchronization poses a fundamental problem for applications dealing with multiple pieces of multimedia signals, such as view synthesis and A/V experience reconstruction.

2.1.1 Synchronization Challenges and Prior Art

In professional video productions such as sports TV broadcasting, the recording cameras may be synchronized based on coordinated hardware and communication protocols to provide synchronized timestamps and ensure accurate temporal alignment. For distributed and ad-hoc settings involving consumer-level devices, different cameras' clocks are not easily synchronized to the frame level. Although an increasing number of devices are now equipped with GPS capabilities that may supply a universal time stamp, many low-cost and low-power devices are not. One line of prior art developed mobile applications and protocols to provide synchronized time tags on video frames [20]. As it requires that users of mobile devices must first install an application and/or follow a prescribed protocol before their videos can be aligned, this approach cannot deal with unconstrained streams for general and less intrusive settings in today's social media and emerging big data paradigms. In absence of proactive synchronization mechanisms, the current solution must rely primarily on visual content and/or sound content, which provides limited alignment resolution on media streams at a rather high computational cost. For these reasons, we have seen limited availability of efficient and effective alignment technologies on audio and video streams, which comes in a sharp contrast with the broad availability and adoption of "stitching" tools for still images.

Efforts have been made to address the video synchronization problem using signal processing and computer vision techniques, as shown in the literatures [13, 14, 46, 54, 56, 63]. In [56], a three-step approach using a set of corresponding fea-

ture points is proposed. The method in [13] simultaneously estimates both spatial alignment and temporal synchronization between two video sequences using spatio-temporal information. In [14], alignment is achieved by matching trajectories of moving objects. The authors in [54] present a synchronization method based on detecting camera flashes present in the video content. Most of the existing work exploits certain features from the video signals that can be detected in one video sequence and related to corresponding features in other sequences. These feature-based methods rely on the visual content of the videos and may not always perform well. For example, it is difficult to synchronize video sequences using visual features when they do not share sufficient common contents of the scene, or the viewing angles differ significantly. Some methods impose constraints on the video recordings to be synchronized, such as calibrated or static cameras and homography between image sequences. These constraints may not always be satisfied in practice.

2.1.2 New Synchronization Modality via ENF

We propose a new modality for video synchronization by exploiting the ENF signal naturally embedded in video recordings [60,61]. In viewing the ENF signal as a continuous-time random process, its realization in each recording may serve as a timing fingerprint. Synchronization of audio and video recordings therefore can be performed by matching and aligning their embedded ENF signals. This approach differs greatly from how the existing work tackles the audio/video synchronization problem, and it has several advantages over conventional methods. The ENF based

method does not rely on the analysis of audio and visual content in the recordings to be synchronized. This property allows it a strong potential to address such difficult scenarios that remain intractable with existing methods. In video synchronization, for example, the conventional approaches based on visual cues do not perform well in situations where arbitrary camera motion occurs or the view overlap is insufficient. The ENF based method is not affected by these adverse conditions. Additionally, extracting and aligning ENF signals may be more effective computationally than the approaches that rely on computer vision and/or extensive learning, so more (or longer) recordings could be efficiently processed. It can also be easily generalized to synchronize multiple pieces of recordings.

ENF signals may be extracted from the soundtracks of video recordings, as well as the image sequences if the video captures the subtle flickering of lightings. Extracting the weak ENF signal from image sequences presents a challenging task. The temporal sampling rate of visual recordings is generally too low for estimating the ENF signal that may appear at harmonics of 50 or 60 Hz. The ENF traces in video signals are relatively weak and may easily be distorted by object and camera motions. In this chapter, we focus on exploiting ENF signals from video sound tracks for synchronization. The extraction of ENF from visual recordings will be addressed in the next chapter.

2.2 Multimedia Synchronization using ENF from Audio

2.2.1 Estimation of ENF from Audio Signal

A general and easily implementable approach to estimating the ENF signal from a source signal, such as audio, is the short-time Fourier transform (STFT), which is a popular non-parametric tool for frequency analysis of time-varying signals. It divides a signal into possibly overlapping frames of small durations. Within every frame, the signal can be regarded as wide-sense stationary, and each of the frames undergoes Fourier analysis respectively. For ENF estimation, we apply STFT to a source signal that contains ENF traces, and find the peak frequency within a certain range near the nominal value or the harmonics in each frame. The values of the peak frequency from all the frames are concatenated to form the estimated ENF signal. Plotting the squared magnitude of an STFT for each time frame, a spectrogram is used to visualize the spectrum of the source signal. The spectrogram is usually displayed as a two-dimensional intensity plot with the two axes being time and frequency, respectively.

Comparisons of various frequency estimation approaches for ENF are carried out in [30, 44]. In [37], the authors evaluate the use of quadratic interpolation over DFT samples as a means to improve the accuracy of ENF estimation. A multi-tone harmonic model is used to build a Maximum-Likelihood estimator for ENF in [12]. Similarly, the use of multiple harmonics is also investigated in [29]. The weighted energy method as used in [23] is adopted here for its robustness and low complexity.

The audio signal is divided into frames of certain length (e.g 8 seconds), and fast Fourier transform (FFT) is calculated for every frame. The ENF signal is then estimated as the dominant frequency within a small vicinity of the nominal value of ENF:

$$F(n) = \frac{\sum_{l=L_1}^{L_2} f(n,l)|s(n,l)|}{\sum_{l=L_1}^{L_2} |s(n,l)|}, \tag{2.1}$$

where f_s and N_{FFT} are the sampling frequency of the signal and the number of FFT points, respectively; $L_1 = \frac{(f_{ENF}-\Delta f)N_{FFT}}{f_s}$ and $L_2 = \frac{(f_{ENF}+\Delta f)N_{FFT}}{f_s}$, where f_{ENF} is the nominal value of ENF and Δf defines the range of frequencies in consideration; $f(n,l)$ and $s(n,l)$ are the frequency and energy in the l^{th} frequency bin of the n^{th} time frame, respectively.

According to the sampling theory, if a signal contains no components of frequency higher than F Hz, it can be completely determined by its sampled version when the sampling rate exceeds $2F$ Hz. The ENF signal embedded in multimedia recordings is usually present close to its nominal value (50 or 60 Hz) and/or several higher order harmonics. Digital audio recordings are usually created with a sampling rate that is much higher than the value of ENF. Therefore, we can down-sample an audio signal before extracting its ENF signal to reduce computational cost.

The ground-truth ENF signal can be obtained from power outlet measurements using a step-down transformer and a voltage divider circuit. Fig. 2.1 shows an example of ENF extraction from audio signal. In this example, an audio recording and a power measurement recording were made simultaneously in the US where the nominal value of ENF is 60 Hz. The ENF signal can be extracted from around

any harmonics of the nominal value of ENF, as long as the ENF traces are strong enough. Here, we examine the second harmonic for the audio recording and the base frequency for the power recording. As can be seen from Fig. 2.1, the ENF signals estimated from the audio recording exhibit variation trends similar to the ground-truth ENF signal from the power outlet measurements.

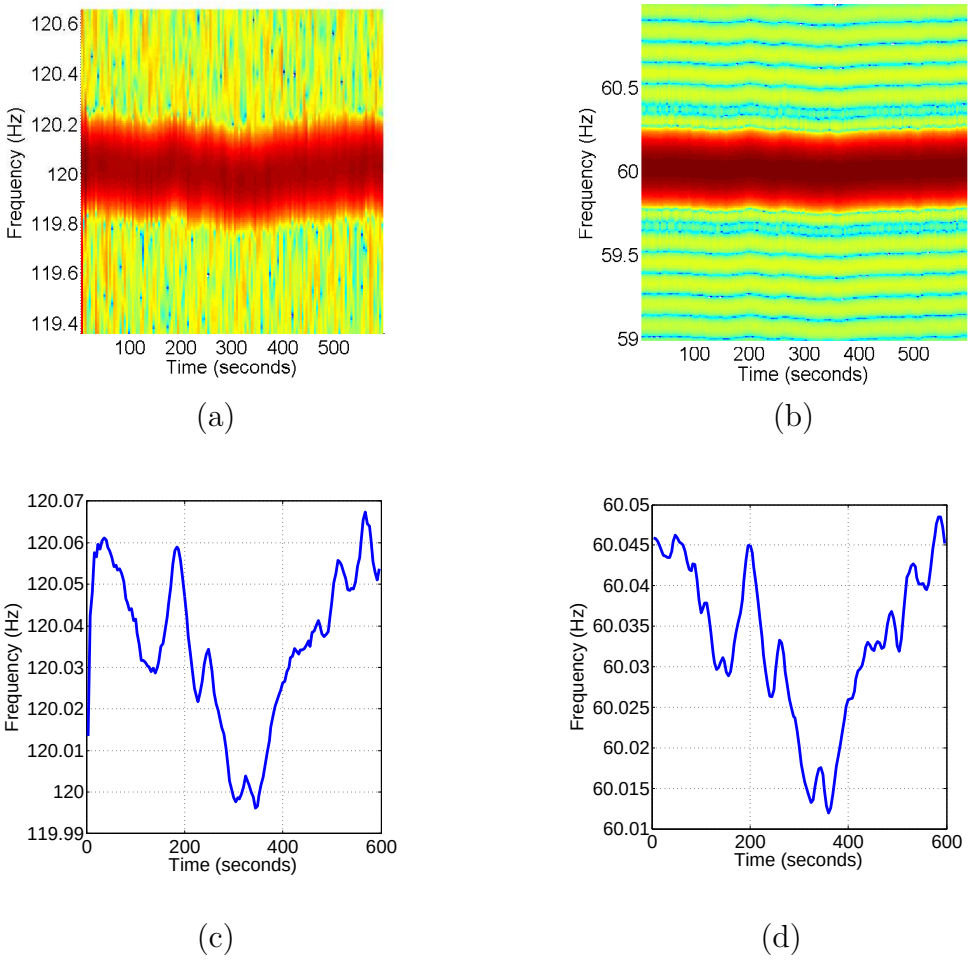


Figure 2.1: Spectrograms and ENF estimates from audio and power signals recorded at the same time. (a) Spectrogram of the test audio signal around the 2^{nd} harmonic (120 Hz); (b) Spectrogram of the power signal around the base frequency (60 Hz); (c) ENF signal estimated from the audio signal; (d) ENF signal estimated from the power signal.

2.2.2 Video Synchronization by Matching ENF Signals

In this section, we discuss in detail how the ENF traces embedded in video soundtracks can be used for video synchronization. After taking the soundtracks from two video recordings to be synchronized, we first divide each soundtrack into overlapping frames of length L_{frame} seconds. The overlap between adjacent frames is denoted as $L_{overlap}$ in seconds. So the shift from one frame to the next is $L_{shift} = L_{frame} - L_{overlap}$. For every frame, we estimate the dominant frequency near the nominal value of the ENF. The values of the estimated frequency are concatenated to form the ENF signal of each soundtrack. The normalized cross correlation coefficients are calculated with different lags between the ENF signals. The lag corresponding to the maximum correlation coefficients is identified as the temporal shift between the two videos. In the following, we show the experimental results of the proposed approach using battery-powered consumer-level cameras.

In the first experiment, we made two video recordings of people playing racquetball in a gym with a Canon PowerShot SX230 HS camera and a Canon PowerShot A2300 camera, respectively. During the recording, the cameras were fixed on tripods shooting the racquetball court from different viewing angles. Both recordings are approximately 10 minutes long, and one begins approximately 20 seconds earlier than the other.

The ENF signals are estimated from the soundtracks of the video clips, and their NCC is calculated with different values of lags between them. In Fig. 2.2 (a), we plot the normalized cross correlation (NCC) as a function of the lag, and observe a

clear peak at 20.52 seconds. We then align the video clips by shifting them relatively by 20.52 seconds. The ENF signals after alignment, along with the reference ENF measured from the power outlet, are shown in Fig. 2.2 (b). Both the ENF signals extracted from the videos exhibit variation trends that are consistent with those of the reference ENF signal. Fig. 2.2 (c) shows a few sample pairs of images from the video sequences after alignment. The images in the same row are from the same video stream, while the images in the same column correspond to the same time instance. By examining the girl's movement in the images, we can see that the two video sequences are well synchronized after using the proposed approach.

For the second experiment, we use the same cameras to make two video recordings in an office building. The cameras were placed in different rooms facing a parking lot through glass windows. We estimate the ENF signals from the soundtracks of the two video recordings and calculate their correlation coefficient as a function of the lags between them, which is plotted in Fig. 2.3 (a). A peak value of 0.95 is found at the lag of 92.16 seconds. The estimated ENF signals after alignment and their reference ENF signal are shown in Fig. 2.3 (b). Although some segments of the ENF signal estimated from one of the video soundtracks (video 1) suffer from some short-term distortions around the 110^{th} second, the video clips are still well synchronized, as shown by the sample image frame pairs in Fig. 2.3 (c).

In the examples shown in this section, the cameras are fixed during recording, and some view overlap occurs between the videos to be synchronized. These settings are chosen so that we can appreciate the effectiveness of the proposed approach by visually examining the image frames. It should be noted that, as an advantage over

many existing methods, no constraints exist on the camera motion or calibration. Also, no requirement is necessary for any view overlap when using the ENF signals extracted from the soundtracks for synchronization of video recordings. Readers may note the substantial differences of view points and depth of the two cameras in both examples. For such situations, even when some view overlap occurs, it can be challenging to handle with traditional computer vision based approaches.

The accuracy of synchronization is important for many applications involving multiple videos. Experiments are conducted to evaluate the synchronization accuracy of the proposed method. We use two audio recorders to make a total of around 17 hours of audio recordings in an office room. A beep sounds at the beginning of the recording so the ground truth of the lag between the recordings can be obtained by comparing the waveforms of the beep using audio editing software. The recordings are divided into clips of L_c seconds, and each clip is treated as an individual test sample. We apply the proposed method to synchronize the test clips and examine the synchronization accuracy under two variables: the length of the test clips L_c, and the strength of the ENF signal in the source signal measured by the signal to noise ratio (SNR). Assume that the ENF signal is estimated near f_0 Hz, which may be the nominal value of ENF or one of its harmonics. The SNR is estimated as the ratio of the energy density in the narrow frequency band $(f_0 - f_{\delta 1}, f_0 + f_{\delta 1})$ to that in the band $(f_0 - f_{\delta 2}, f_0 - f_{\delta 1})$ and $(f_0 + f_{\delta 1}, f_0 + f_{\delta 2})$, where $f_{\delta 2} > f_{\delta 1}$. To test the more challenging scenarios with low SNR levels, white Gaussian noise of controlled strength is added to the original source signals.

Fig. 2.4 shows the synchronization error of each test clip pair and the estimated

SNR of the ENF signal. As expected, the synchronization accuracy improves with higher SNR and longer duration of clip. The average absolute synchronization errors for the test clips without added noise are 0.48, 0.40, and 0.17 seconds for clip sizes of 5, 10, and 15 minutes, respectively.

Although most demonstrations of ENF being picked up by digital audio and video recordings in areas of electrical activities have been reported in the recent decade, the presence of ENF can be found in analog recordings made throughout the second half of the 20^{th} century. In our recent work, we demonstrated that ENF traces can be found in digitized versions of 1960s phone conversation recordings of President Kennedy in the White House [58]. Using ENF to analyze historical recordings could offer many useful applications for forensics and archivists. For instance, many 20^{th} century recordings are important cultural heritage records, but lack necessary metadata, such as the date and time of recording. Also, the need may arise to timestamp old recordings for investigative purposes, and ENF may provide a solution.

As an example of exploring historical recording synchronization using the ENF signals, we analyze two recordings from the 1970 NASA Apollo 13 mission [1] that we know were recorded at approximately the same time. The first recording comes from the PAO (Public Affairs Officer) loop, which is the space-to-ground communications that was broadcast to the media. The second recording is of GOSS Net 1 (Ground Operational Support System), which is the recording of the space-to-ground audio as the people in mission control heard it. Both recordings are approximately four hours long. Figure 2.5 shows spectrogram strips for both recordings about the

ENF harmonics. We observe that for the first recording, the ENF clearly appears around all the harmonics, and it is especially strong around 360Hz. For the second recording, the ENF is noisier, and it appears best around 120Hz and 360Hz.

We extract the ENF of the first recording from around 360Hz. For the second recording, we use the spectrum combining technique for ENF estimation [29], where we combine the ENF traces from around 120Hz and 360Hz to arrive at a more reliable ENF estimate. The resulting ENF signal remains rather noisy; we clean the signal by locating outliers and replacing them using linear interpolation from surrounding ENF values. Figure 2.6 (a) shows 20-minute simultaneous ENF segments from both recordings, with the second ENF signal displaced by 0.05Hz to distinguish them and see them separately. Visually, the two signals appear similar.

In a synchronization scenario, we would need to match ENF segments from two or more signals with potentially different lags, then decide on the correct lag based on the similarity of the segments, using the correlation coefficient as a metric. As a proof-of-concept for the Apollo data described above, we divide the first Apollo ENF signal into overlapping 10-min ENF segments, and for each segment, we correlate it with equally-sized segments from the second Apollo ENF with varying lags. The two signals were recorded at the same time, so the ground truth suggests that the highest correlation should be at zero lag. Figure 2.6 (b) shows the mean values of the correlations achieved for different lags, and we can clearly see that the highest correlation is achieved for zero lag, which matches the ground truth.

We observe that the techniques discussed earlier for audio and video alignment can be extended to aligning two historical recordings of interest. This can potentially

help timestamp old recordings of unknown date of capturing. With old recordings, we may not always have access to reference power ENF, as in the case considered here, yet we have the potential to utilize historical recordings of known date and time to create an ENF database to which we can compare recordings of interest with uncertain information about capturing time.

2.2.3 Discussions

The prerequisite for the proposed approach is that ENF signals can be extracted well from the audio recordings to be synchronized. This prerequisite may not always be satisfied in practice. For example, audio recordings created by battery-powered recorders at locations far from electrical equipments are likely not to capture ENF traces. We have conducted experiments to study the existence of ENF signals. We use battery-powered audio recorders to make recordings at multiple locations inside a university building. These locations include hallways, classrooms, offices, computer labs, and study lounges. The reference power recordings are recorded as well. With a total of 100 test audio recordings, we are able to extract ENF signals that match with the power reference from 81 of them. It proves that most of audio recordings in the building can capture ENF signals.

Statistical modeling and analysis of ENF signals can be performed to study ENF signal's capability of synchronization theoretically. It is shown in [25] that the ENF signal can be modeled as an autoregressive random process. Based on this model, a decorrelation-based approach is proposed to improve the matching

accuracy. Such study can help understand the effect of ENF signal's duration and signal-to-noise ratio to the synchronization accuracy.

2.3 Chapter Summary

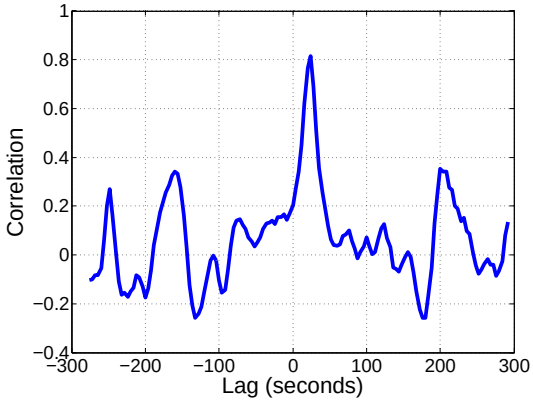
In this chapter, we propose to synchronize audio and video recordings by extracting and matching the ENF signal that is naturally embedded in the audio clip during its creation. The value of the ENF signal fluctuates around its nominal value randomly over time. The variation patterns of the ENF are very similar within the same power grid, even at distant locations. Audio recordings can often capture the ENF signal due to electromagnetic interference to the recording sensor or the electric humming. By correlating the ENF signals extracted from audio recordings or the soundtracks of videos, multiple audio and video clips can be synchronized.

We have conducted experiments to test the synchronization accuracy of the proposed approach. With a fair signal-to-noise ratio for the ENF signal embedded in the audio signal, the average synchronization error can be as low as about 0.12 second, which is equivalent of 3-4 frames for most video clips created by consumer cameras with a frame rate at 30 fps.

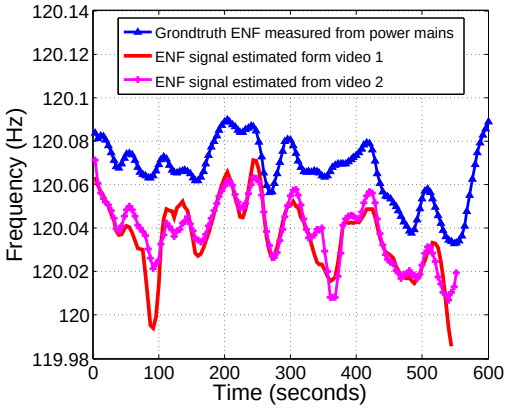
Our study also shows that the ENF signal may be present in historical recordings as well. In particular, we are able to extract ENF signal from some digitized analog recordings of the NASA Apollo 13 Mission. As an example, we have successfully synchronized two audio clips that were recorded at roughly the same time but at different loops. Such analysis of old recordings may be of great value for revealing

unknown historical facts and archiving historical events.

The traces of ENF may also be found in visual recordings, such as the image sequence of video clips. Extracting the weak ENF signal from image sequences presents a challenging task. The temporal sampling rate of visual recordings is generally too low for estimating the ENF signal that may appear at harmonics of 50 or 60 Hz. The ENF traces in video signals are relatively weak, and may be easily distorted by object and camera motions. In the next chapter, we focus on the extraction of ENF from visual recordings.



(a)



(b)

(c)

Figure 2.2: Example-1 of video synchronization by aligning the ENF signals from soundtracks. (a) The correlation coefficients of the ENF signals as a function of the relative lag between them. (b) The ENF signals from the two video recordings after alignment and the ENF measured from the power mains at the same time. (c) Several sample frame pairs after alignment. Rows correspond to video sequences, and columns correspond to time instances.

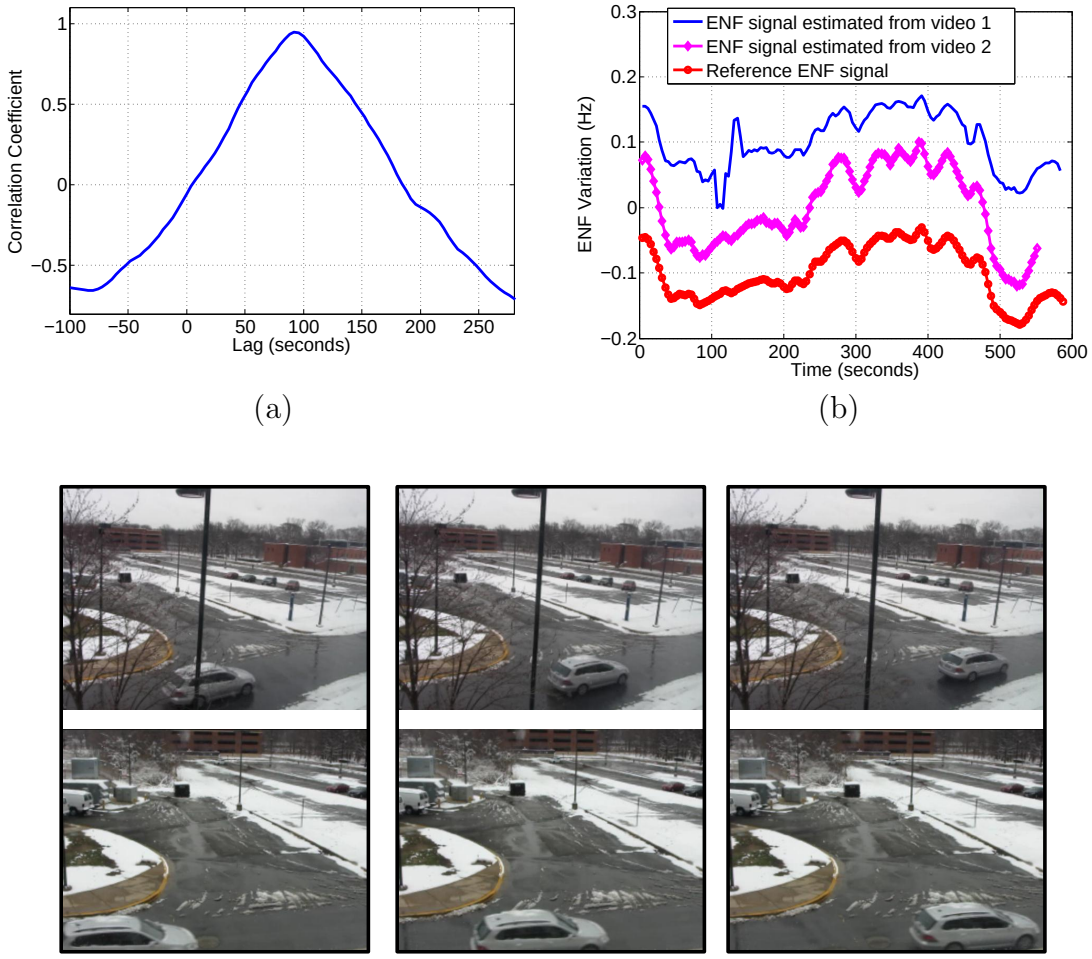
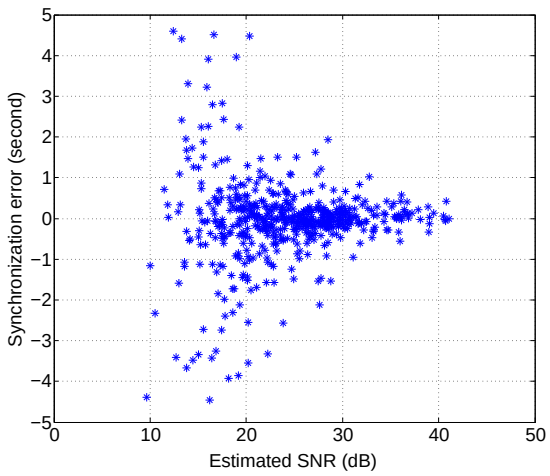
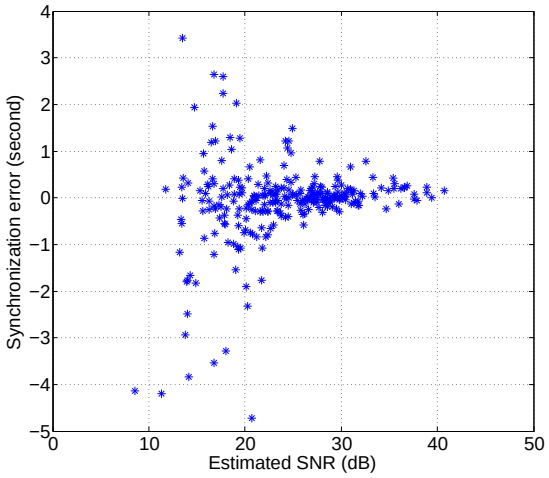


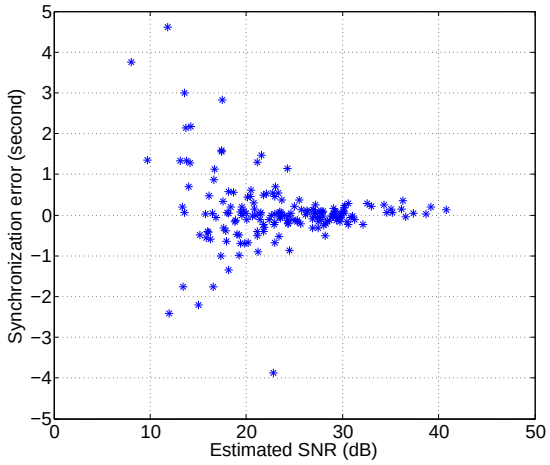
Figure 2.3: Example-2 of video synchronization by aligning the ENF signals from soundtracks. (a) The correlation coefficients of the ENF signals as a function of the relative lag between them. (b) The ENF signals from the two video recordings from different viewing angles after alignment and the ENF measured from the power mains at the same time. The signals are properly shifted to facilitate comparison. (c) Several sample frame pairs after alignment. Rows correspond to video sequences, and columns correspond to time instances.



(a)

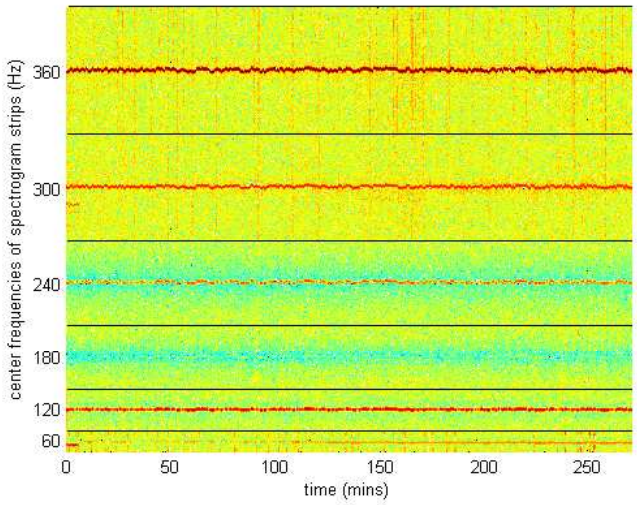


(b)

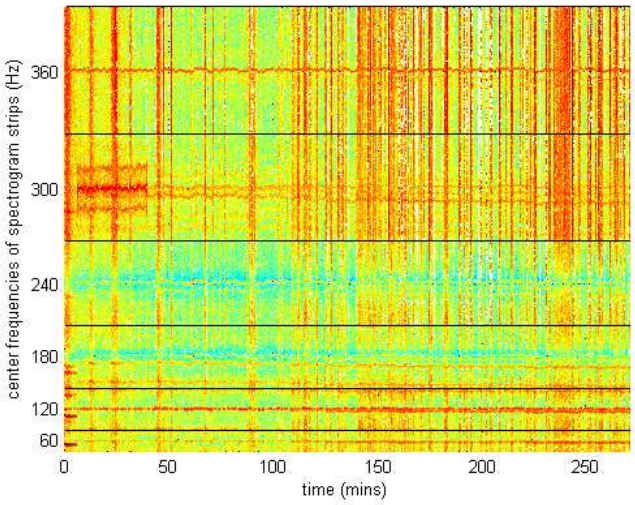


(c)

Figure 2.4: The synchronization error with clip size being (a) 300 seconds; (b) 600 seconds; (c) 900 seconds.

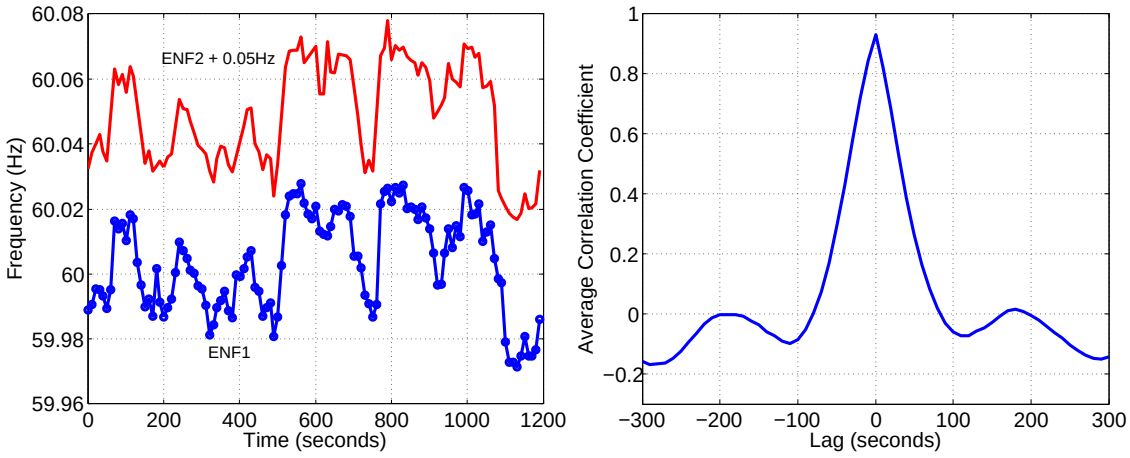


(a)



(b)

Figure 2.5: Spectrogram strips around the ENF harmonics for the Apollo 13 recordings. (a): PAO recording; (b): GOSS recording.



(a) (b)

Figure 2.6: Synchronize the Apollo 13 mission recordings with the ENF signals.

Chapter 3

ENF Extraction from Visual Recording

3.1 Chapter Introduction

Most previous work related to the analysis of ENF signals is built on extracting ENF traces from audio recordings [27, 29, 50, 51]. Recently, it has been found that indoor lightings such as the fluorescent lights and incandescent bulbs vary their light intensity in accordance with the AC supply voltage, which varies according to the AC supply frequency [23]. As a result, cameras can capture the light intensity variation that can be used to extract the ENF signal. In [23], the authors took the mean of the pixel values in every frame of video recordings that capture indoor lightings, then used spectrogram analysis to estimate the embedded ENF signal. The aliasing effect presents a major challenge of this scheme. Most of the consumer-level digital cameras adopt a frame rate of around 30 fps, while the ENF signal appears at the harmonics of 50 or 60 Hz. Therefore, the ENF signal suffers from a severe

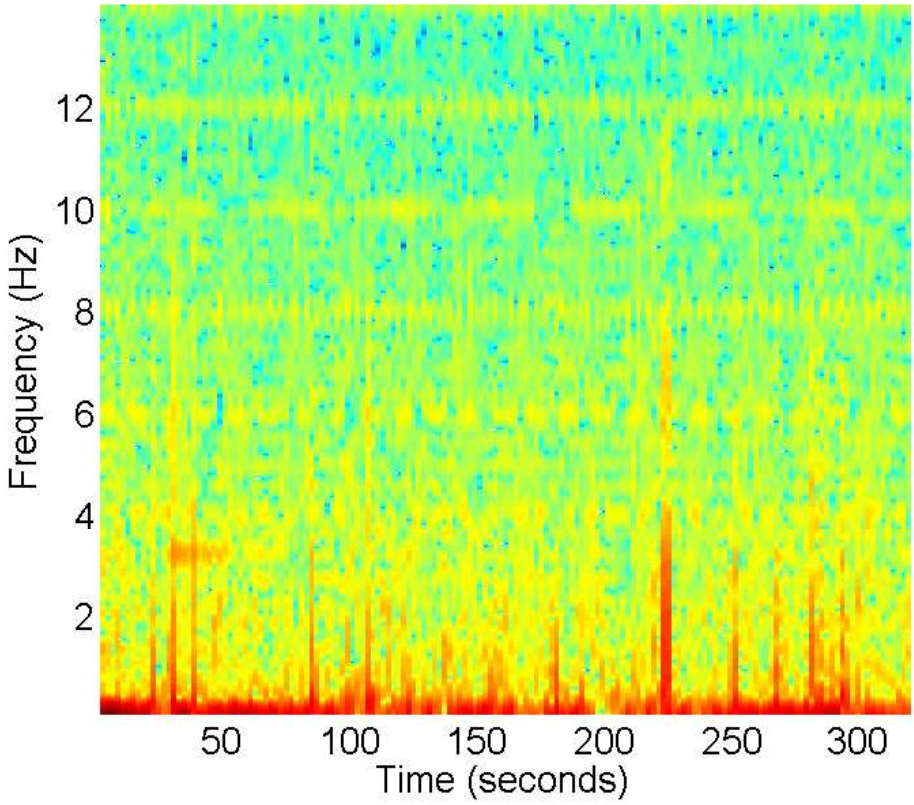


Figure 3.1: The spectrogram of the mean values of the frames from a test video recording shooting a white wall under under fluorescent lighting. The ENF signal overlaps with the DC components and is difficult to extract. Figure is best viewed in color.

aliasing effect induced by insufficient sampling speed. In the US, the nominal value of the ENF is 60 Hz. If the frame rate is exactly 30 Hz, the ENF signal will be shifted to 0 Hz, i.e., the DC frequency. As a result, it is highly difficult to estimate the ENF signal due to low signal-to-noise ratio. Figure 3.1 shows the spectrogram calculated from the mean values of the frames of a video shooting a white wall under fluorescent lighting. We observe the ENF signal overlaps with the DC components and is difficult to extract.

It has been proposed to take advantage of the rolling shutter to address the problem of insufficient sampling rate [24]. Rolling shutters are commonly adopted for the complementary metal-oxide semiconductor (CMOS) camera sensors. Cameras with global shutters in charge-coupled device (CCD) sensors record an entire frame from a snapshot of a single point in time. However, a camera with a rolling shutter scans the vertical or horizontal lines of each frame sequentially, so that different lines in the same frame are exposed at different times. If we treat each line of the frame as a sample, the temporal sampling rate can be much higher than the frame rate, which would facilitate the estimation of the ENF signal.

In this work, we conduct a further study on the exploitation of the rolling shutter for extracting ENF traces from video recordings. We model and analyze the rolling shutter mechanism with a filter bank, then perform analysis using multirate signal processing theory. We extend the scope of ENF extraction from videos of still scenes to those containing motions, which presents a more challenging problem and has never been formally attempted. Several methods are developed, and promising results are observed.

3.2 Extracting ENF from Visual Recording

3.2.1 Exploiting the Rolling Shutter Mechanism

With a rolling shutter, each frame is recorded by scanning across the frame either vertically or horizontally line by line, instead of capturing the whole frame at a single point in time as in the case of a global shutter. Figure 3.2 illustrates the

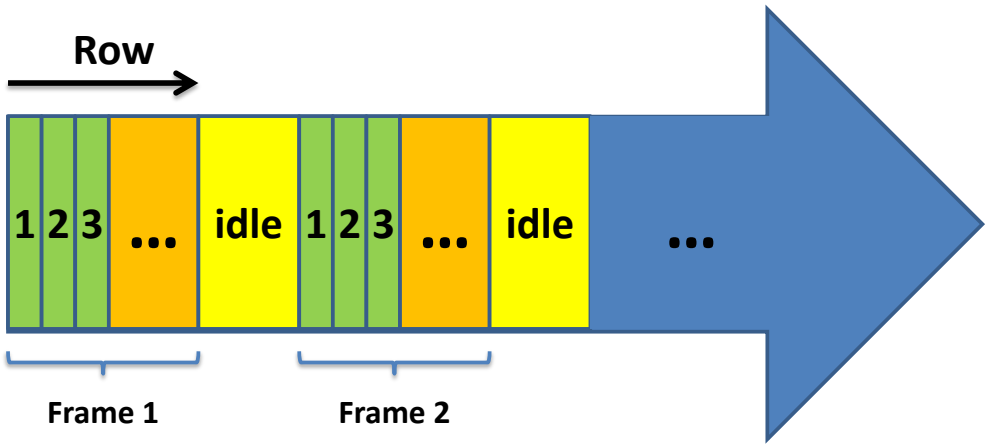


Figure 3.2: Timing of rolling shutter sampling: the rows of a frame are sequentially exposed, followed by an idle period before proceeding to the next frame.

timing for the image acquisition of rolling shutters, assuming the frame scanning is done row-by-row. Each row of the frame is exposed sequentially to light, followed by a possible idle period before proceeding to the next frame. Since pixels in different rows are exposed at different times, but are displayed simultaneously during playback, the rolling shutter may cause such distortions as skew, smear, and other image artifacts, especially with fast-moving objects and rapid flashes of light [35].

The sequential read-out mechanism of a rolling shutter has been traditionally considered detrimental to image/video quality due to its accompanying artifacts. However, recent work has shown that the rolling shutter can be exploited with computer vision and computational photography techniques [7,19,28]. Our previous work [24] exploited the rolling shutter to extract ENF traces from videos of static scenes. In this work, we investigate the more challenging cases of videos containing motions [59].

A Filter Bank Model

For an image captured by a rolling shutter, we can treat the spatial mean of every row as a temporal sample since all the pixels in a single row are exposed at the same time. As a between-frame idle period occurs, in terms of capturing the ENF signal over time, we are equivalently abandoning some samples that would have been generated in the idle period. The time domain illustration of this model is shown in Figure 3.3. Here, we assume that the shutter is able to produce M samples at its full capacity, and only L samples among them are retained while the rest are discarded, where $L \leq M$. We denote the input and output signal as $x(n)$ and $y(n)$, respectively.

To facilitate frequency domain analysis, we use a L-branch filter bank to model the relationship between the input signal $x(n)$ and the output signal $y(n)$, as shown in Figure 3.4. In each branch of the filter bank, the input goes through an M-fold down-sampler followed by an L-fold up-sampler, with appropriate delays at both the beginning and the end of the branch.

The DTFT of the signal coming from the l^{th} branch can be analyzed according to multi-rate signal processing theory [64]:

$$Y_l(\omega) = \frac{1}{M}\left(\sum_{m=0}^{M-1} X(\frac{\omega L + 2\pi m}{M}) e^{j\frac{\omega L + 2\pi m}{M} l}\right) e^{-j\omega l}. \tag{3.1}$$

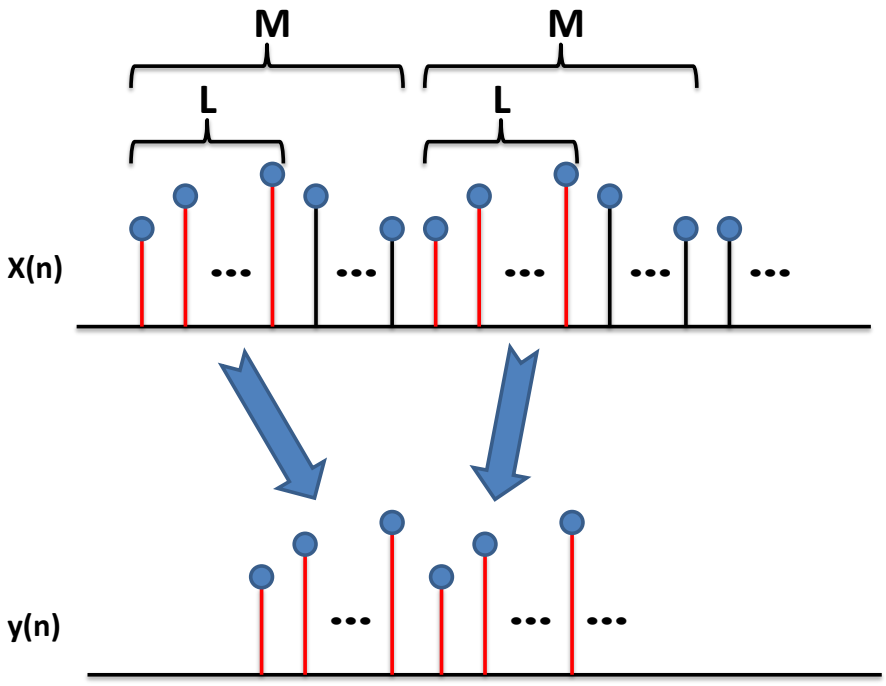


Figure 3.3: A time domain illustration of the rolling shutter sample acquisition. Every L out of M samples are retained, and the other samples that correspond to the idle periods are discareded.

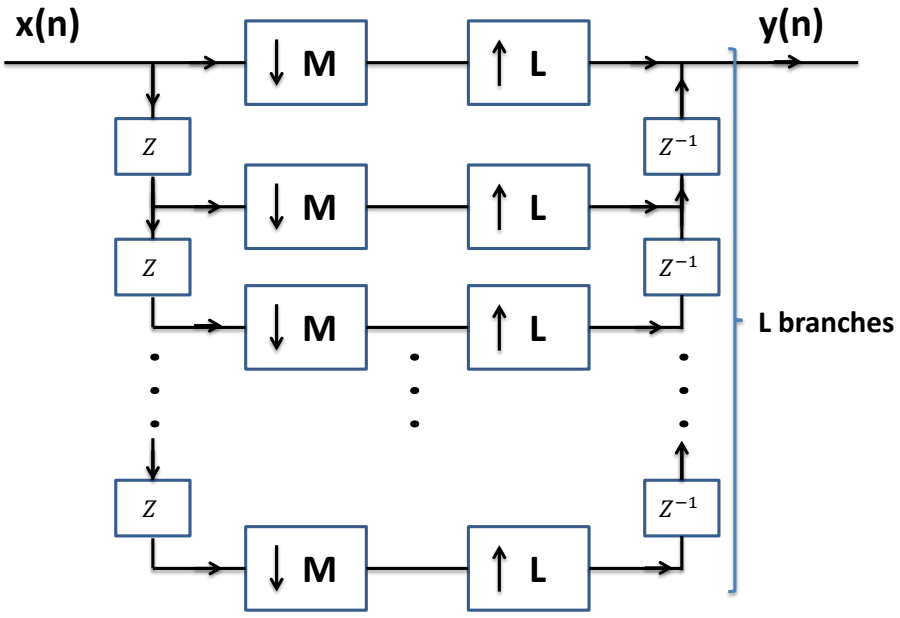


Figure 3.4: A filter bank model of rolling shutter sample acquisition. In each branch of the filter bank, the input goes through an M-fold down-sampler followed by an L-fold up-sampler, with appropriate delays at both the beginning and the end of the branch.

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

$$\begin{aligned} Y(\omega) &= \sum_{l=0}^{L-1} Y_l(\omega) \\ &= \sum_{l=0}^{L-1} \frac{1}{M} \left(\sum_{m=0}^{M-1} X(\frac{\omega L + 2\pi m}{M}) e^{j\frac{\omega L + 2\pi m}{M} l} \right) e^{-j\omega l} \\ &= \sum_{m=0}^{M-1} X(\frac{\omega L + 2\pi m}{M}) F_m(\omega), \end{aligned} \tag{3.2}$$

where

$$F_m(\omega) = \frac{1}{M} \sum_{l=0}^{L-1} e^{-j\frac{\omega(M-L)-2\pi m}{M} l}. \tag{3.3}$$

Equivalently, with $\omega = 2\pi f$, we have

$$Y(f) = \sum_{m=0}^{M-1} X(\frac{\omega L + 2\pi m}{M}) F_m(f). \tag{3.4}$$

T denotes the frame duration of the camera. According to the notations in the previous section, the sampling rate of the shutter is $f_s = M/T$, and the perceptual sampling rate of the row signal is L/T. Here, L is the number of rows per frame, and the exact value of M depends on the CMOS manufacture's design and is usually unavailable to the public. The spectrogram of the row signal is computed using the perceptual sampling rate L/T, instead of the actual sampling rate M/T. Deriving from Equation 3.4, we can show that the Fourier representation of the row signal is

$$Y(f) = \sum_{m=0}^{M-1} X(\frac{2\pi}{f_s}(f + \frac{m}{T})) F_m(f). \tag{3.5}$$

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

3.2.2 ENF Estimation

In this section, we describe how to extract ENF traces from videos captured by rolling shutters. Without loss of generality, we assume the rolling shutter scans the frame row-by-row. Consider a video signal $s(r, c, n)$, where $1 \leq r \leq R, 1 \leq c \leq C$ and $1 \leq n \leq N$ denote the row index, column index and frame index, respectively. The video signal contains mainly two components: one is the visual component v corresponding to the visual scene; and the other is the ENF component e:

$$s(r, c, n) = v(r, c, n) + e(r, c, n). \tag{3.6}$$

The authors in [24] use the spatial average of each row in the video as the source signal to estimate the ENF signal. From Eq. (3.6), we see that the signal-to-noise-ratio (SNR) of e in s may be low in the presence of the visual component v. For fixed spatial indices r and c, the visual component $v(r, c, n)$ as a function of n is in general a low-pass signal. In order to suppress the effect of v and extract the ENF component e, we propose to apply high-pass filtering to the video signal s. In the next sections, we describe the high-pass filtering techniques for two cases: for videos of static scenes and for videos with motions.

We first consider a video recording with a static scene, so the visual signals of every video frame are identical, i.e., $v(r, c, n) = v(r, c)$. Under this assumption, Eq. (3.6) is reduced to

$$s(r, c, n) = v(r, c) + e(r, c, n). \tag{3.7}$$

We can apply a high-pass filter to s by subtracting from it its mean value across all

frames:

$$\begin{aligned}\hat{s}(r,c,n) &= s(r,c,n) - \bar{s}_n(r,c) \\ &= s(r,c,n) - \frac{1}{N}\sum_{m=1}^{N} s(r,c,m) \\ &= e(r,c,n) - \frac{1}{N}\sum_{m=1}^{N} e(r,c,m). \end{aligned} \tag{3.8}$$

Here $e(r,c,n)$ is the sinusoidal ENF signal sampled at the r^{th} row and c^{th} column in the n^{th} frame. For any given r and c, $e(r,c,n)$ as a function of $n = 1, 2, ..., N$ is essentially a sinusoid sampled at the frame rate of the video recording. Since the frequency of the ENF signal is changing over time, $e(r,c,n)$ for $n = 1, 2, ..., N$ tends to have different phases and cancel out. So, for a sufficiently large N, the average of these samples is close to 0, i.e.,

$$\bar{e}_n(r,c) = \frac{1}{N}\sum_{m=1}^{N} e(r,c,m) \simeq 0. \tag{3.9}$$

This leads to

$$\hat{s}(r,c,n) \simeq e(r,c,n). \tag{3.10}$$

After the high-pass filtering, the SNR of the ENF signal in $\hat{s}$ is much higher than that in the original video signal s. We then use the spatial average of each row in $\hat{s}(r,c,n)$ as the source signal to estimate the ENF signal:

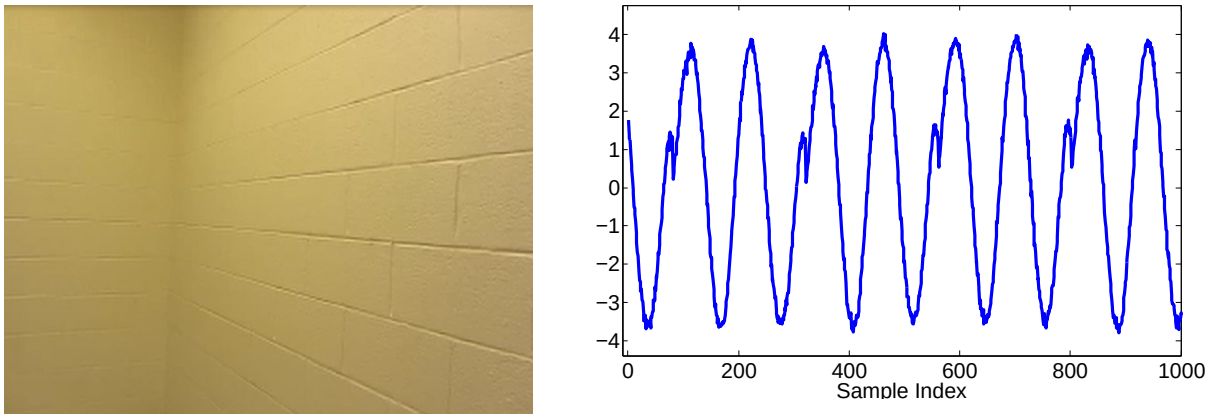
$$R(r,n) = \frac{1}{C}\sum_{c=1}^{C} \hat{s}(r,c,n). \tag{3.11}$$

$R(r,n)$ is referred to as the *row signal* hereafter. We can use the frequency estimation techniques discussed earlier to estimate the ENF signal from the row signal.

We have conducted experiments using a Canon PowerShot SX230 HS camera that is battery-powered and equipped with a rolling shutter. Fig. 3.5 shows an example of ENF estimation from a static video recording. The test video recorded a white wall under fluorescent lighting, with a camera mounted on a tripod. The spatial resolution of each video frame is 480×320, and the frame rate is 29.97 fps. Fig. 3.5 (a) shows a snapshot of the test video. We calculated the row signal according to Eq. (3.11), and then vectorized it by concatenating its entries frame after frame to form the source signal for ENF estimation. Fig. 3.5 (b) shows a segment of the source signal. We observe that the source signal exhibits sinusoidal waveforms, except for some periodic phase shifts. These phase shifts exist because of the idle period of the rolling shutter between exposing the last row of one frame and starting the first row of the next frame. No recording is conducted during the idle period, and a phase jump of the source signal may occur on every frame border (every 240 samples in this experiment).

In this example, the nominal value of ENF is 60 Hz, and the frame rate of the camera is $\frac{1}{T} = 29.97$ fps. The intensity variations of the fluorescent lightings should follow the instantaneous energy of the AC power supply, exhibiting a oscillation of around 120 Hz. By Eq. (3.5), the ENF traces embedded by the row signal from the video recording should appear at around $120 + m \times 29.97$ Hz, where $m = 1, 2, 3, ...$ This holds consistent with what we observe from the spectrogram of the row signal in Figure 3.6.

The ENF traces can be extracted from the spectrogram of the row signals around 30 Hz, 60 Hz, 90 Hz... We estimate the ENF signal from close to 60 Hz as



(a) (b)

Figure 3.5: (a) A snapshot of a test video of a static scene of a white wall. (b) The source signal for ENF estimation.

we see from the spectrogram that the SNR of the ENF signal appears the highest in this frequency range. The estimated ENF signal from this recording along with the reference ENF signal simultaneously measured from the power mains are plotted in Figure 3.7 (a). They are appropriately shifted to lie within the same dynamic range, as only the variation trends are of interest. The ENF signals from the video recording and the power measurement exhibit similar variations. The correlation coefficient between them as a function of the relative time lag is plotted in Figure 3.7 (b), and a clear peak is observed at the ground truth lag of 0 second.

In the second example, the camera was placed in a room illuminated by fluorescent lights, which is representative of real-life surveillance scenarios where few events are expected to occur. Measurements from the power mains at the time of the video recording were also conducted to provide the reference ENF signal. In Figure 3.8 (a), we plot the estimated ENF signals from the video and the power

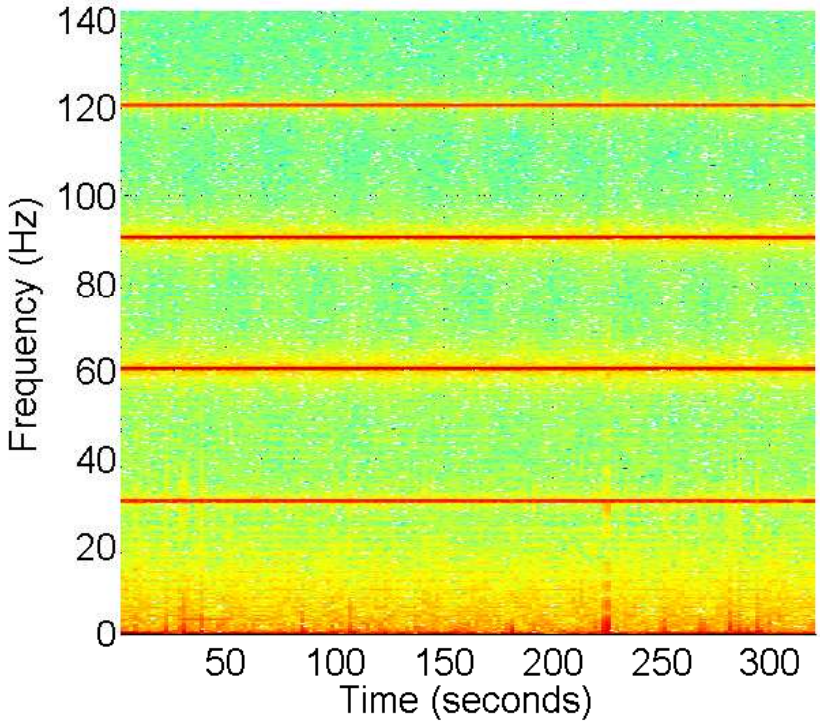
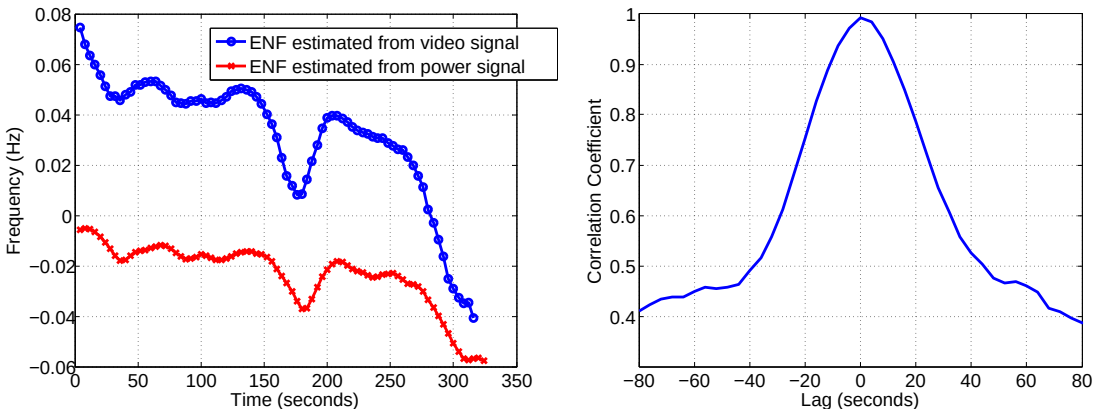


Figure 3.6: The spectrogram of the row signal from a white wall video recording. Figure is best viewed in color.

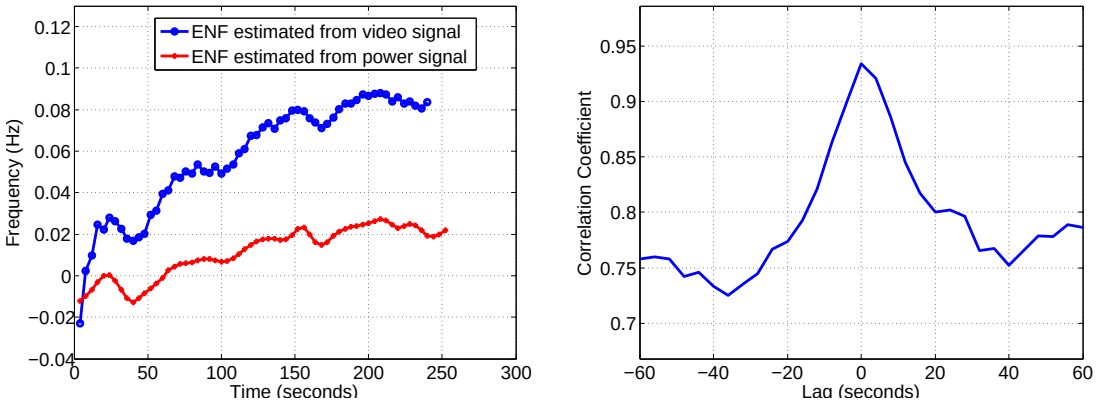


(a) (b)

Figure 3.7: (a) The ENF signals (appropriately shifted) extracted from a test video of white wall and the power measurement. (b) Correlation coefficient between the video and power ENF signal as a function of relative time lag.

measurement, and the correlation coefficient between them as a function of lag is plotted in Figure 3.8 (b).

In the third example, we show that videos of outdoor scenes can also capture ENF traces. A test video is captured in a dark parking lot illuminated by outdoor lighting. The camera is fixed, and no object motion occurs in the recording. Fig. 3.9 (a) shows a snapshot of the test video. The ENF signal estimated from the test video exhibits similar variation patterns to the one estimated from the simultaneous power measurements, as can be observed in Fig. 3.9 (b). It proves that the outdoor lighting are connected to the mains, and the ENF signal is successfully extracted from the video that captures the subtle flickering in the light.



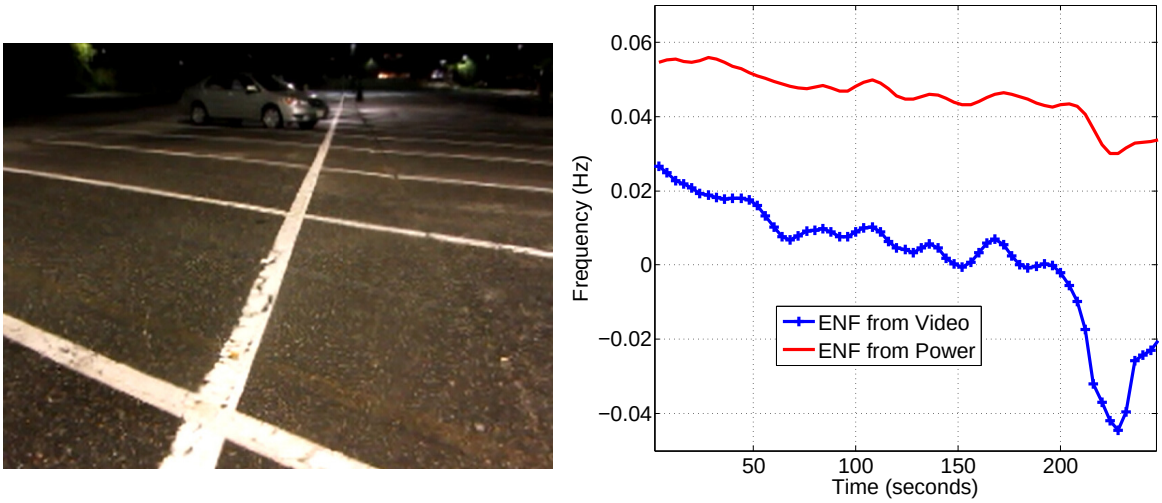
(a) (b)

Figure 3.8: (a) The ENF signals (appropriately shifted) extracted from a surveillance video and the power measurement. (b) Correlation coefficient as a function of relative time lag.

3.2.3 Videos with Motions

Extracting ENF signals from video recordings of scenes with moving objects proves to be more challenging. In such scenario, Eq. (3.7) does not hold anymore, and the method for high-pass filtering in the previous subsection would no longer work.

If the scene in the video contains a static background, we can use the static regions to estimate the ENF signal. Following the notations of last subsection, given two image frames $s(r, c, n)$ and $s(r, c, m)$, we are interested in finding the regions not affected by object motion in either frame. The mutual motion-free regions between



(a) (b)

Figure 3.9: (a) Screen shot of the test video of a parking lot. (b) The ENF signals (appropriately shifted) extracted from the test video and the power measurement.

$s(r, c, n)$ and $s(r, c, m)$ are represented by a binary matrix $M^{n,m}(r, c)$, defined as

$$M^{n,m}(r,c) = \begin{cases} 1 & \text{if frame } n \text{ and frame } m \text{ are both static at pixel (r,c)} \\ 0 & \text{otherwise} \end{cases}$$

As an example, Fig. 3.10 (a) and (b) show two sample images from a video sequence recorded inside an office building. The motion-free regions shown in Fig. 3.10 (c) are found by thresholding the pixel-wise differences of the pixel intensity between the two images.

Using a strategy similar to that in Sec. 3.2.2, we apply a high-pass filter to the video signal by subtracting from it a smoothened version of the original signal. For an image frame s from the video sequence, we search for its mutual motion-free regions against all the other frames. The pixel values of the frames in their respective motion-free regions can be averaged to form a smooth version of s, which

(a) (b) (c)

Figure 3.10: (a) Frame 1. (b) Frame 2. (c) Mutual motion-free regions are highlighted.

is then subtracted from s:

$$\hat{s}(r,c,n) = s(r,c,n) - \frac{1}{\sum_{m \neq n} M^{n,m}(r,c)} \sum_{m \neq n} s(r,c,m) \cdot M^{n,m}(r,c) \tag{3.12}$$

The row signal is obtained by taking the row average of $\hat{s}$, from which the ENF signal can be estimated. We have conducted an experiment with a video that records people walking in an office building's hallway. The video used similar settings as the experiments in Sec. 3.2.2. We use the proposed scheme to extract the ENF signal from this test video. The reference ENF signal is also estimated from a simultaneously recorded power signal. We observe from Fig. 3.11 that the variation trends of the ENF signal estimated from the test video remain consistent with those of the reference ENF signal.

In a second example, we made a video recording of a moving Hexbug toy in a room with indoor lighting. The Hexbug is a robotic toy with fast movement, powered by the vibrations of a built-in battery motor. Figure 3.12 shows several

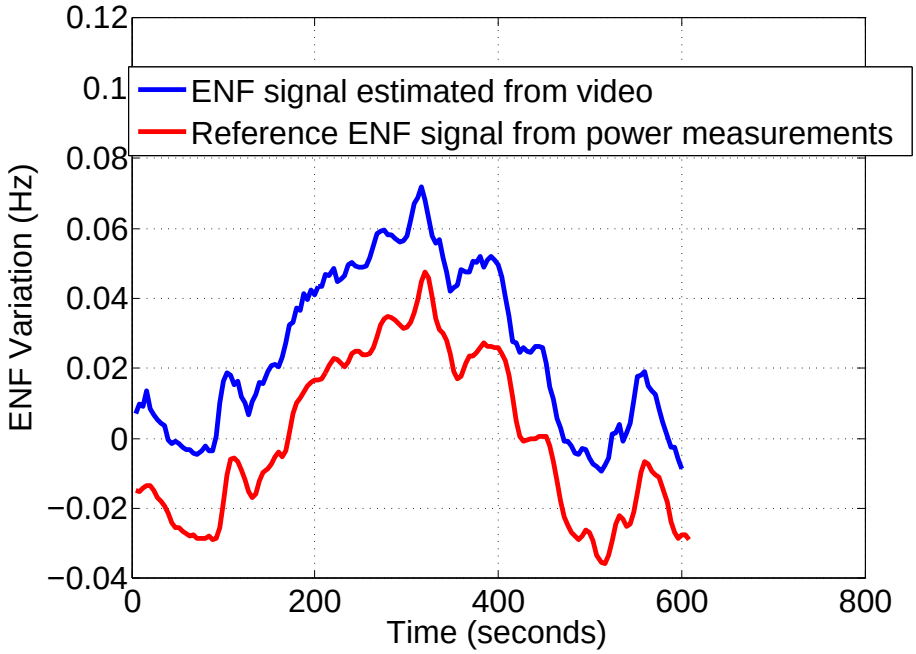


Figure 3.11: The ENF signal estimated from the test video matches well with the reference ENF signal. The signals are properly shifted to facilitate comparison.

video frames. The ENF signals extracted from the Hexbug video and its reference ENF signal from power mains are plotted in Figure 3.13 (a). They exhibit similar variations, and a correlation peak is observed when they align, as seen from Figure 3.13 (b).

3.2.4 Brightness Change Compensation

Many cameras have a brightness control mechanism that adjusts the camera's sensitivity to light in response to the illumination conditions, so the overall brightness of an acquired image remains visually pleasing. As an example, two images from a video sequence are shown in Fig. 3.14. As the person in the second image is closer to the camera, the background wall appears brighter than in the first image.

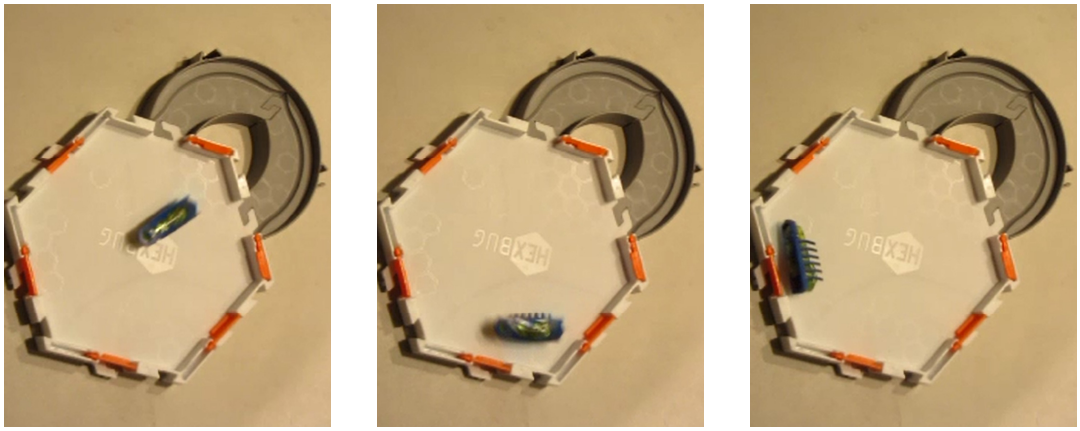
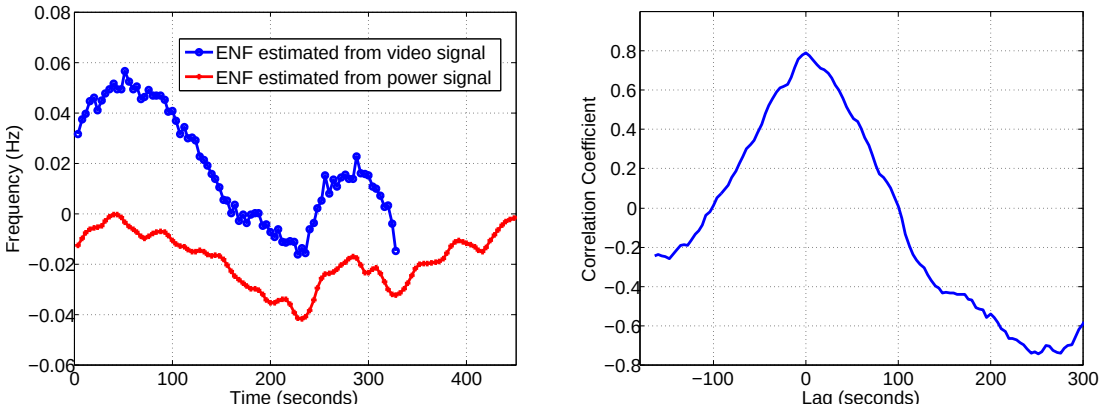


Figure 3.12: Sample frames of the hexbug test video.

Such brightness changes introduce challenges to the estimation of the ENF signal when using the techniques described in previous subsections.

To investigate how to mitigate he negative effect of brightness changes, we have created the following recording: during the first four minutes, a person walked around in a hallway relatively distant from the camera so the camera's automatic brightness adjustment was not triggered; after four minutes, the person moved closer to the camera and brightness changes occurred, as shown in Fig. 3.14. The ENF signal is extracted from this test video using the techniques discussed in previous subsections without addressing the brightness changes. In Fig. 3.16, we observe that the estimated ENF signal from the test video becomes distorted after four minutes into the recording as a result of brightness changes.

We have examined the relationship of the pixel values in different images of the same scene. For two images, we find the regions in which both images remain static. Each pixel in the static regions is represented as a dot in Fig. 3.15, whose X coordinate is the pixel value in image 1, and the Y coordinate is the pixel value in



(a) (b)

Figure 3.13: (a) The ENF signals extracted from the test video of the Hexbug and the power measurement (appropriately shifted). (b) Correlation coefficient as a function of relative time lag.

image 2. From the figure we observe that the brightness change can be well modeled by a linear transform. Given two frames $s(r, c, n)$ and $s(r, c, m)$, we have

$$s(r, c, n) = a^{n,m} \cdot s(r, c, m) + b^{n,m}. \tag{3.13}$$

For a frame $s(r, c, n)$, the pixel values in the static background regions are used to estimate the parameters $a^{n,m}$ and $b^{n,m}$. For brightness change compensation, we apply Eq. (3.13) to each frame $s(r, c, m)$. Eq. (3.8) then becomes

$$\begin{aligned}\hat{s}(r, c, n) =& s(r, c, n) - \frac{1}{\sum_{m \neq n} M^{n,m}(r, c)} \cdot \\ & \sum_{m \neq n} \left(a^{n,m} \cdot s\left(r, c, m\right) + b^{n,m}\right) \cdot M^{n,m}(r, c)\end{aligned} \tag{3.14}$$

The described scheme was applied to the test video, and the result of ENF estimation is shown in Fig. 3.16. With our proposed brightness change compensation, the ENF signal estimated from the test video now exhibits consistent variations with

Figure 3.14: Two image frames from a test video recording demonstrating camera's automatic brightness control mechanism.

the reference ENF signal.

3.2.5 Camera Motion Compensation

In previous discussions, we have assumed that the camera is fixed during recording so that the pixels in different image frames align spatially. In practice, people may hold the camera by hand to make a video recording, so the camera would undergo certain movement and the previously described methods may not apply.

In order to address the situations with camera motions, we first consider videos of static scenes. For two image frames $s(r, c, n)$ and $s(r, c, m)$, we denote by $(\delta_r^{n,m}, \delta_c^{n,m})$ the pixel-wise shift between them due to the camera motion:

$$s(r, c, n) = s(r + \delta_r^{n,m}, c + \delta_c^{n,m}, m). \tag{3.15}$$

To compensate for the camera motion, we need to shift the pixels in two frames relatively by $(\delta_r^{n,m}, \delta_c^{n,m})$ so that they align spatially. The registered frames can be

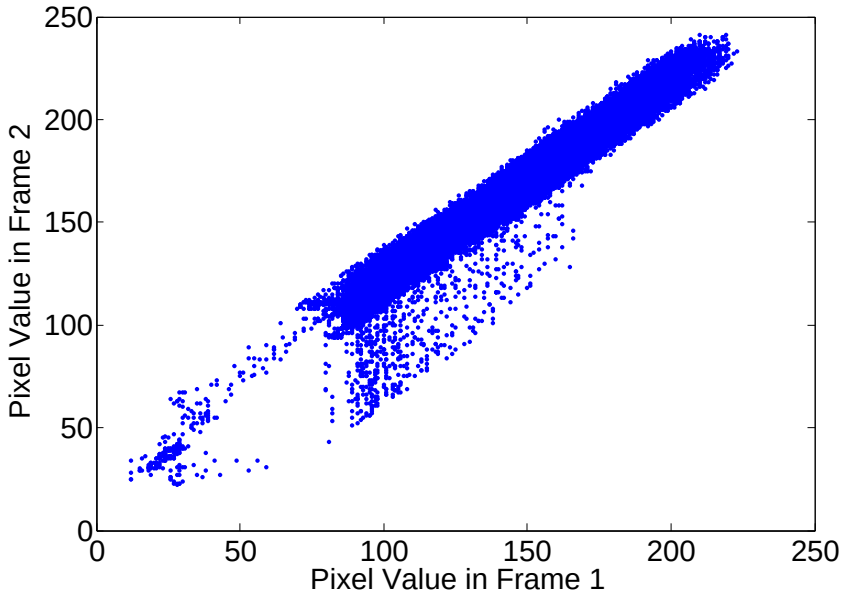


Figure 3.15: The relationship between the pixel values in static regions of two images subject to brightness change.

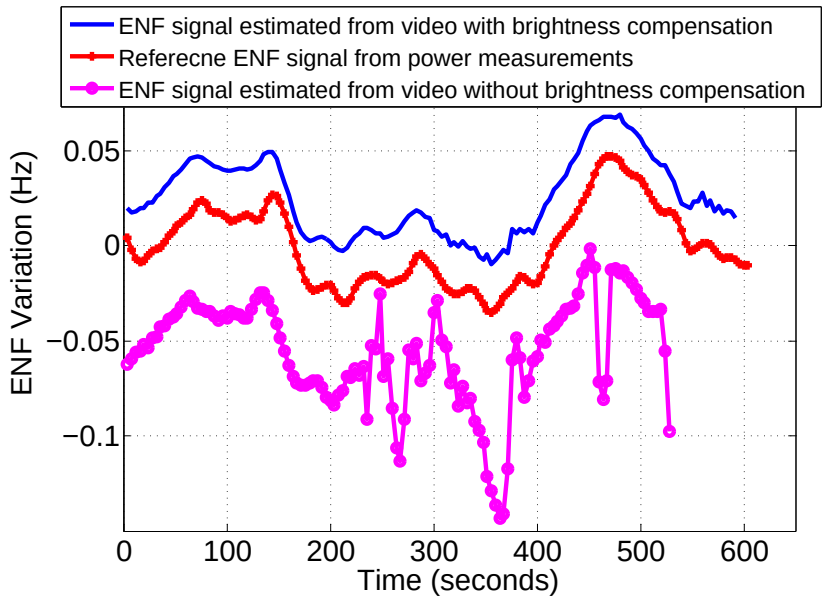


Figure 3.16: The effectiveness of the brightness change compensation technique. The signals are properly shifted to facilitate comparison.

processed as described in the previous subsections. Considering the camera motion compensation, Eq. (3.8) becomes

$$\hat{s}(r,c,n) = s(r,c,n) - \frac{1}{N}\sum_{m=1}^{N} s(r+\delta_r^{n,m}, c+\delta_c^{n,m}, m), \tag{3.16}$$

and the ENF signal can be estimated from $\hat{s}(r,c,n)$.

Optical flow methods can be used to estimate the pixel-wise displacement between image frames. These methods calculate the motion field (V_r, V_c) between two frames $s(r,c,n)$ and $s(r,c,n+\delta_n)$ based on the optical flow equation $\frac{\partial s}{\partial r}V_r + \frac{\partial s}{\partial c}V_c + \frac{\partial s}{\partial n} = 0$, and certain additional conditions and constraints for regularization. In this work, we have used the implementation of the optical flow estimation developed by [36].

An experiment was conducted to verify the proposed camera motion compensation scheme. We used the Canon PowerShot SX230 HS camera to record a video of a hallway. The camera was hand-held during the recording, and we deliberately manipulated the camera to create noticeable motion, as shown in the sample frames in Fig. 3.17. If we ignore the camera motion, the spectrogram of the source signal extracted from the test video exhibits blurry spectral energy distributions, as shown in Fig. 3.18 (a). The ENF signal estimated from the test video without camera motion compensation is shown in Fig. 3.18 (b), and it deviates from the reference ENF signal. We then apply the proposed camera motion compensation to the test video. The ENF signal estimated from the video signal after compensation matches well with the reference ENF signal as shown in Fig. 3.18 (b).

If object motion occurs in the scene in addition to camera motion, we can

Figure 3.17: Sample frames from a test video with camera motion.

apply camera motion compensation first to the image frames, then use the strategy in Sec. 3.2.3 to locate and utilize the static backgrounds for extraction of ENF signals.

3.3 Video Synchronization by Matching ENF from Image Sequence

In Chapter 2, we demonstrated video synchronization by aligning the ENF signals extracted from the soundtracks of video clips. In certain scenarios, such as some surveillance recordings, video recordings may have been muted or soundtracks may have been edited, and thus have no reliable audio available. As an alternative,

we may extract the ENF signal from the image sequence of the visual track using the techniques described in Sec. 3.2. In this section, we present experimental results of this approach.

We used two Canon PowerShot SX230 HS cameras equipped with CMOS sensor and rolling shutter to video an office building's hallway illuminated by an indoor light. The cameras were placed to capture the hallway from different view angles, and each recording is about eight minutes long. A person walked through the hallway back and forth, and his movements were captured by both cameras.

We apply the methods discussed in Sec. 3.2 to estimate the ENF signals from the image sequences of both video recordings. The NCC of the estimated ENF signals as a function of the lags between them is plotted in Fig. 3.19 (a), from which we find a peak NCC value of 0.96 at 60.72 seconds. The ENF signals after alignment are shown in Fig. 3.19 (b), and we observe the variation patterns of the ENF signals match well. In Fig. 3.19 (c), we show several image frames from the synchronized video recordings. For comparison, we manually align the two videos by comparing the image frames and the soundtracks in both video clips, and found the lag to be 60.80 seconds, which is close to the value obtained by the proposed approach.

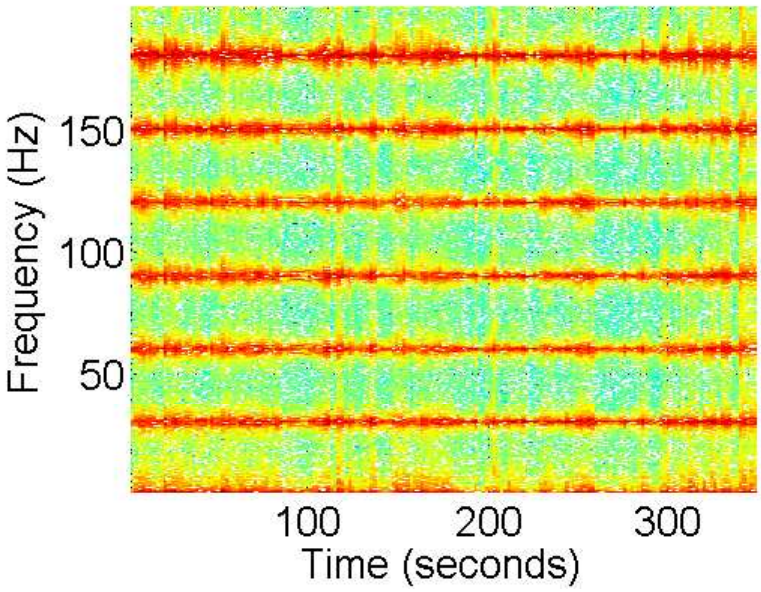
3.4 Chapter Summary

In this chapter, we have studied the problem of extracting ENF traces from videos. We have exploited the rolling shutter of CMOS imaging sensors and treated each line as an ENF impacted signal sample in order to compensate for the low

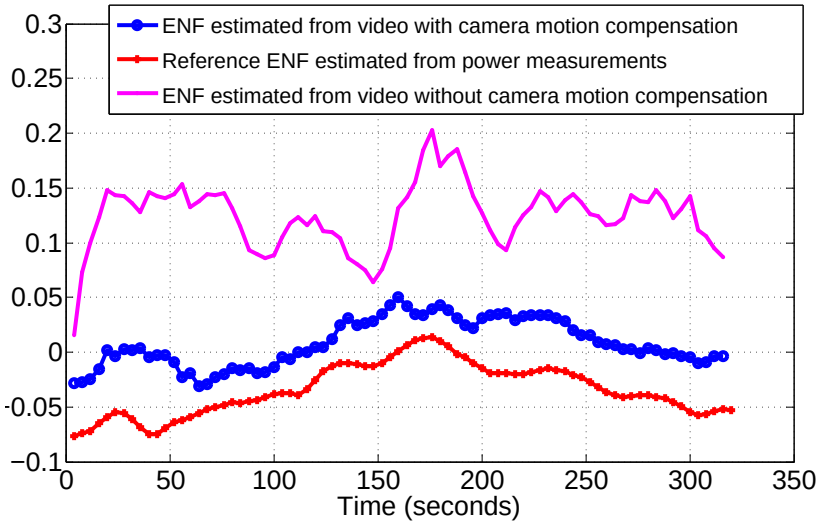
frame rate of video recordings. The rolling shutter mechanism was analyzed using a filter bank model and multirate signal processing theory. Extraction of the ENF signal from an image sequence present many challenges, and, to the best of our knowledge, few research attempts have been made to adequately address it. We have proposed several techniques to overcome the difficulties of extracting the ENF signal from image sequences, such as the low sampling rate, object motions in the scene, camera motions, and brightness change. Through our experiments, we have demonstrated that video recordings can be accurately synchronized by aligning the inherently embedded ENF signals.

In this chapter together with Chapter 2, We have demonstrated promising results of video synchronization using the naturally embedded ENF signals in the soundtracks and image sequences of video clips. As a prerequisite for this method, the ENF traces available in the audio/video recordings must be strong enough for reliable estimation. Through our current study and experiments, we find that this prerequisite may not always be satisfied. The recordings may have been created in an environment where the ENF traces are relatively weak. For example, audio recordings created by recorders powered by batteries at locations far from electrical equipment, and videos made in areas without any electric lighting would probably not capture any ENF traces. Also, the embedded ENF signals may suffer distortions. Most audio and video recorders apply compression to the output recordings, and some strong compression artifacts may adversely affect the estimation of ENF signals. For audio recordings made in loud environments, the strong foreground sounds or voices may also bring distortions to the ENF signals.

Most conventional methods of audio/video synchronization extract and match certain audio and visual features from the contents of recordings. The approach proposed in the paper does not rely on the audio or visual contents of the multimedia signals, and therefore is fundamentally different from and complementary to existing work. The ENF-based approach and the audio/visual-cue-based approach may complement each other and may be combined to solve the problem of multimedia synchronization more effectively.

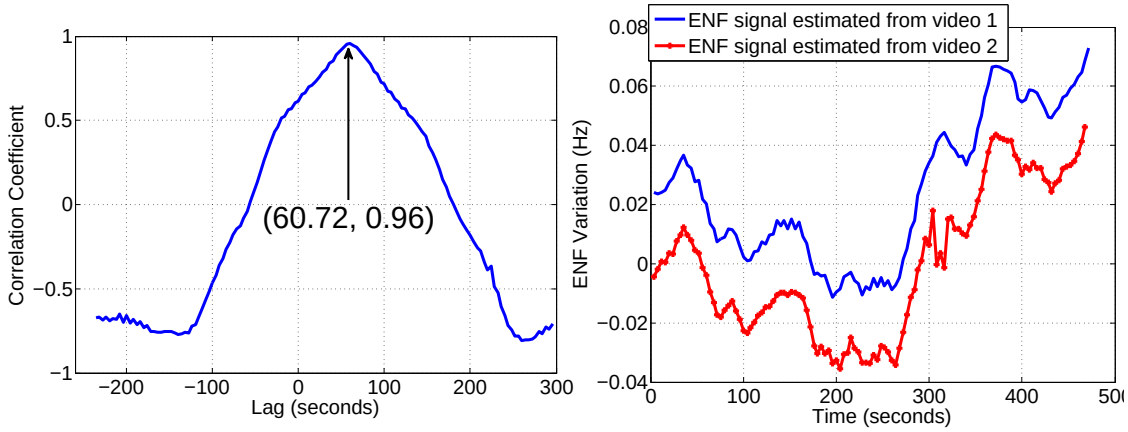


(a)

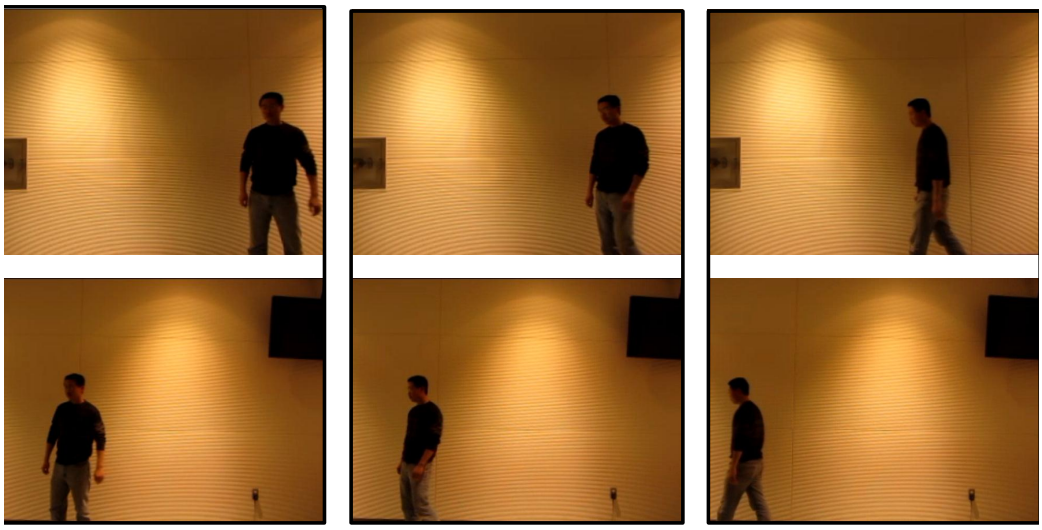


(b)

Figure 3.18: Experiment of a video recording with camera motion. (a) The spectrogram of the source signal without camera motion compensation. (b) The ENF signals estimated from the test video. The signals are properly shifted to facilitate comparison.



(a) (b)



(c)

Figure 3.19: Example of video synchronization by aligning the ENF signals. (a) The correlation coefficient of the ENF signals as a function of the relative lag between them. (b) The ENF signals estimated from the two video recordings after alignment. The signals are properly shifted to facilitate comparison. (c) Several sample frame pairs after alignment. Rows correspond to video sequences, and columns correspond to time instances.

Chapter 4

ENF Analysis on Historical Audio Recordings

4.1 Chapter Introduction

ENF signal is embedded into multimedia recordings at the time of recording, which leads to several interesting questions about the ENF traces in recaptured audio recordings. If recapturing of recording happens in the region of the same nominal ENF as the original recording, the ENF traces, due to the two recording processes, may overlap with each other. How will such overlap affect the quality of the ENF signal extraction? ENF signals in recaptured audio recordings may contain two components: one is inherited from the original recording, referred to as the *content ENF* signal; the other is embedded during recapturing process, referred to as the *original ENF* signal. The original and recapturing ENF signals may have different energies; the signal with a higher energy is referred to as the *dominant ENF*

and the one with a lower energy as the *latent ENF*. The ENF signals in a recaptured audio recording can provide useful information about the audio file in question, such as when the original recording was created and when it was recaptured.

The question of ENF extraction in recaptured audio is relevant to analyzing recordings of historical importance. For example, such historical recordings as NASA Apollo lunar mission audio recordings [1, 2] and President Kennedy's White House conversations [9] were conducted in the analog era of 1960's. These recordings were recently digitized and made available online. Several interesting tasks can be accomplished using such recordings. For example, multiple channels of NASA Apollo mission recordings can be used to create a time synchronized exhibit of the mission. As an ENF signal is time-varying, it can potentially be used to align multiple audio recordings archived from such historical events. However, due to the digitization process, the recordings available online may also have been affected by the ENF signals corresponding to the time of digitization. To the best of our knowledge, no prior work has addressed the effect of recapturing of audio recordings on ENF signals.

As will be shown in this chapter, conventional ENF estimation techniques can only extract the dominant ENF signal. This observation motivates us to design algorithms to extract both the dominant and the latent ENF signals from recaptured recordings. Audio recapturing can also be used as an "anti-forensic" strategy by an adversary to alter ENF traces to mislead a forensic examiner, so developing techniques to extract multiple overlapping ENF signals may also complement the existing techniques to counter such anti-forensic operations [17]. In this work, we

propose a decorrelation based algorithm to estimate both the dominant and the latent ENF from a recaptured audio. After estimating the dominant ENF using conventional ENF signal estimation techniques, a residual signal is computed by subtracting the estimated dominant ENF signal from the original signal. The latent ENF is then estimated from the residual signal.

Besides the possible superposition of multiple ENF traces, another distortion of the ENF signals that is unique to digitized analog tape recordings is the "drifting effect". Due to mechanical imperfection of the analog recorders and tapes, the rolling speed of these tapes often varies during recording and replay. The inconsistency between the rolling speeds of original recording and playback during digitization induces speed errors in the digitized audio. For example, if the tape rolls faster during digitization than during the original recording, the digitized version of the recording will play faster than normal. As a result, the ENF signal in a digitized analog audio recording may deviate from its original value. In this work, techniques have been proposed to compensate for such drifting effect to achieve accurate ENF analysis. We also demonstrate that ENF signals suffering such drifting effect can be exploited to detect and correct tape speed errors.

4.2 Distortions of ENF Signal in Historical Audio Recordings

4.2.1 Multiple ENF Traces in Recaptured Audio Recordings

As this work aims to analyze the ENF signals present in a recaptured audio, we conduct the following experiment to test the robustness of ENF signals to

recapturing. An audio is recorded in an office using a digital recorder. To simulate the conditions of recapturing, we play this recording on a stand-alone speaker in an acoustic anechoic chamber and record the playback using a digital recorder. Fig. 4.1(a) and (b) show the spectrograms of the original recording and the recaptured recording, respectively. From these figures, we observe that the ENF signal is present at the harmonic frequency of 240 Hz in the original recording and the recaptured recording. High correlation is observed between the ENF signals extracted from the original and recaptured recordings around 240 Hz. When we switch-off the replayed audio, the energy peak present at 240 Hz in the spectrogram of the recaptured audio recording disappears. This happens because no interference is present from power lines at 240 Hz in the acoustic chamber. In this example, the content ENF signals and the recapturing ENF signals do not interfere with each other.

In another example, Fig 4.2 (a) shows the spectrogram of a historical recording from President Kennedy's White House conversations, which are available at [9]. This recording occurred in 1962 on analog tape and was digitized later. From this figure, we observe that two different ENF signals are present near 240 Hz, and one of them (present around 239 Hz) disappears well before the end of audio. After listening to the audio, we note that the original recording is turned off at this time. We conjecture that the 239 Hz signal is the original ENF signal and the 240 Hz signal is the recapturing ENF signal.

The two examples have demonstrated the case when the original ENF signal and the recapturing ENF signal are non-overlapping. From such recordings, both the ENF signal can be extracted easily by using suitable bandpass filters followed

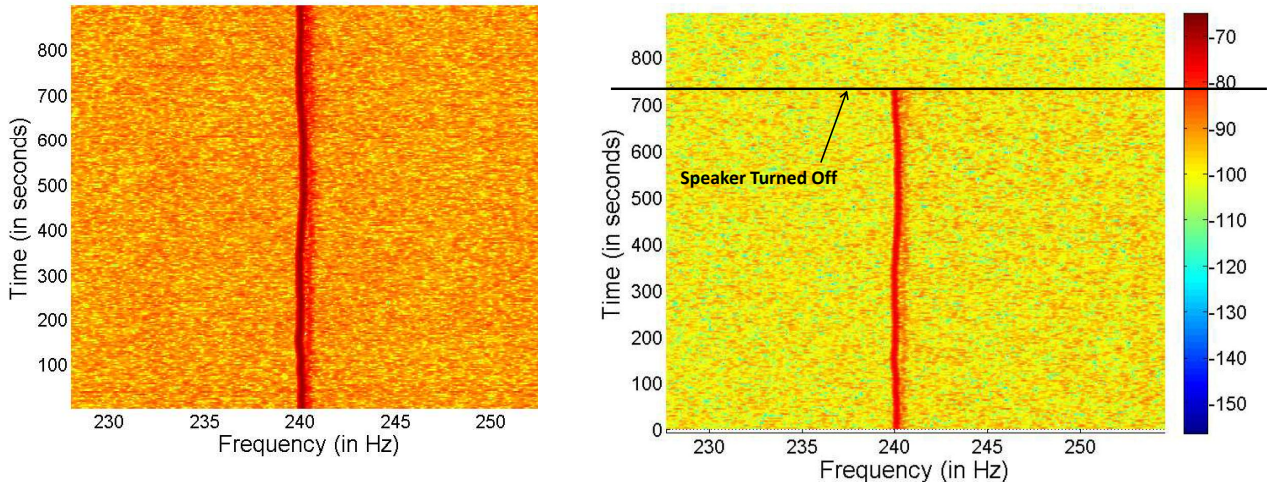
by conventional ENF estimation techniques around the frequency of interest. In less favorable cases, however, the original ENF signal and the recapturing ENF signal may overlap and interfere with each other. To illustrate this scenario, we conduct a recording in the acoustic chamber and recapture it in the same place. As the ENF signal in the same room is embedded from the electromagnetic influences of the same power sources, the original ENF and the recapturing ENF are overlapping at a frequency of 120 Hz. From the spectrogram of the recaptured audio shown in Fig. 4.2 (b), we observe that for the duration of the playback of the original audio on the speaker, the original ENF and the recapturing ENF overlap with each other and the energy distribution of the spectrogram appears noisy. After the original audio is switched-off, the ENF signal becomes cleaner as only the recapturing ENF is captured.

Conventional ENF signal estimation methods extract dominant frequencies in a narrowband around the frequency of interest (nominal ENF or its harmonics) of a given signal. As the original ENF signal and the recapturing ENF signal in recaptured recordings may overlap with each other at the same frequency range, the conventional methods fail to extract both the ENF signals. To demonstrate this, we generate two frequency sequences, $E_d(t)$ and $E_l(t)$, as following:

$$E_d(t) = 60 + N_d(t)$$

$$E_l(t) = 60 + N_l(t),$$

where $N_d(t)$ and $N_l(t)$ are drawn from i.i.d. Gaussian random processes of zero mean and variance 0.1. Using these two signals, we generate a time domain signal



(a) (b)

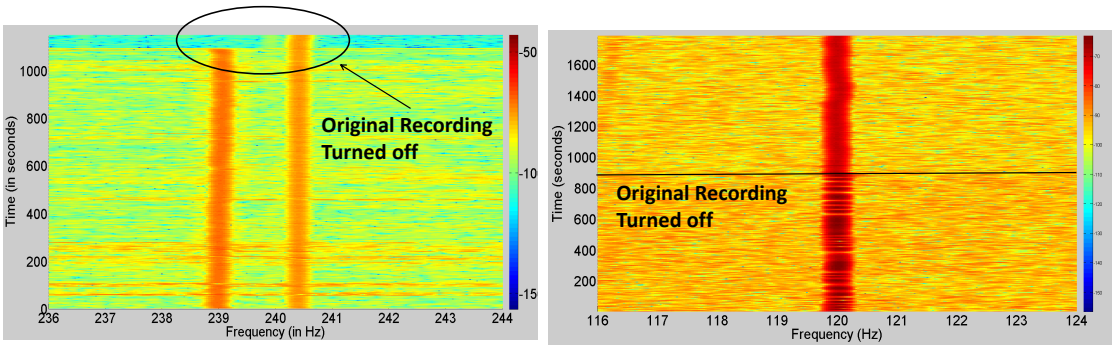
Figure 4.1: Spectrograms of the original and recaptured recordings. (a) original audio; (b) recaptured audio

that varies according to the frequencies $E_d(t)$ and $E_l(t)$ as follows:

$$\begin{aligned} s(t) &= \cos(2\pi \int_0^t E_d(\tau)d\tau) + \\ &\quad \sqrt{\alpha}\cos(2\pi \int_0^t E_l(\tau)d\tau) + N(t), \end{aligned} \tag{4.1}$$

where $N(t)$ is a Gaussian random process of zero mean and unity variance, and α is a constant with $0 \leq \alpha \leq 1$. We see from Eq. 4.1 that signal $s(t)$ consists of two sinusoids of different amplitudes with $E_d(t)$ and $E_l(t)$ being their instantaneous frequencies at time t. Based on this model of $s(t)$, $E_d(t)$ is the dominant ENF signal and $E_l(t)$ is the latent ENF signal, as the energy of the sinusoid corresponding to $E_d(t)$ is greater than $E_l(t)$. This model of signal $s(t)$ is similar to when the original ENF signal and the recapturing ENF signal overlap.

We use a weighted energy frequency estimation method to extract the ENF



(a) (b)

Figure 4.2: Audio recording spectrograms. (a) digitized Kennedy conversation recording; (b) recaptured audio signal with overlapping ENF signals.

signal from $s(t)$. The spectrogram is computed for $s(t)$, and the ENF for each time bin is estimated by weighing frequency bins around the nominal ENF value (60Hz) with the energy present in the corresponding frequency bins. We compute the normalized cross-correlation (NCC) between the estimated ENF signal and the ground truth frequency sequences $E_d(t)$ and $E_l(t)$, respectively. The experiment is repeated multiple times with different realizations of $N_d(t)$, $N_l(t)$, and $N(t)$. The mean and the variance of the NCC values obtained for different values of α is shown in Fig. 4.3. From this figure, we observe that when there is a significant difference between the energy of the dominant ENF and the latent ENF, the correlation between the extracted ENF and the dominant ENF is high (0.6-0.7 range). However, as the energy of the latent ENF signal increases, this correlation value decreases and becomes low (< 0.3 for α close to 1). Similar results were obtained for other frequency estimations methods, such as the subspace based MUltiple SIgnal Classification (MUSIC) and Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT).

Our preliminary results also show that these subspace-based approaches can only obtain reliable estimates when a sufficient margin between $E_d(t)$ and $E_l(t)$ exists. This experiment on synthetic data verifies that the conventional ENF estimation techniques fail to extract the overlapped ENF signals, which is usually the case with recaptured audio recordings. In the following subsection, we describe a new algorithm to extract both the dominant and the latent ENF from recaptured audio recordings.

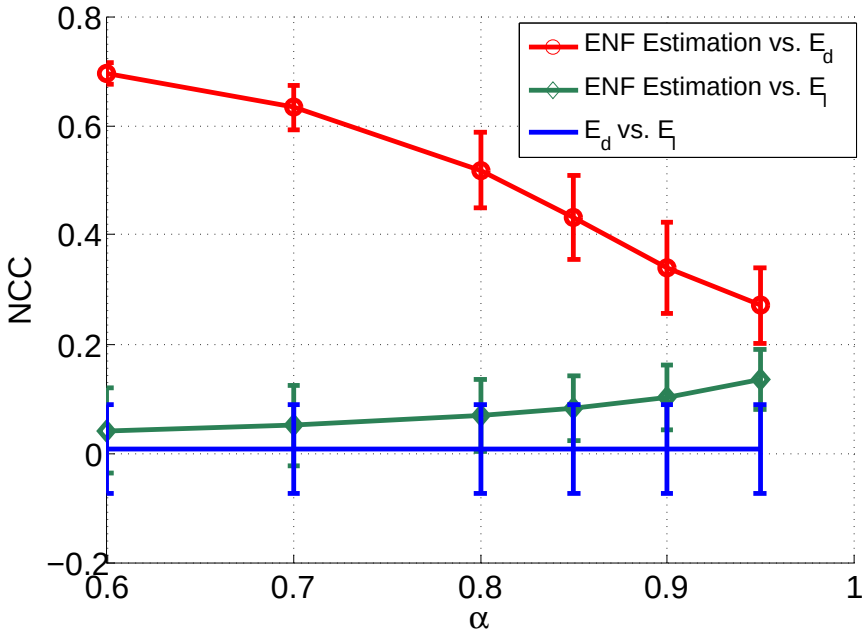


Figure 4.3: The mean and the variance of the NCC values.

A Decorrelation-Based Solution

Our proposed algorithm to extract both the dominant and the latent ENF signals in a recaptured audio works in two stages: first the dominant ENF is estimated, followed by the latent one. As discussed earlier, the dominant ENF, denoted

by $E_d(t)$, in an audio s(t) can be estimated using conventional estimation techniques such as the weighted energy method. After estimating $E_d(t)$, we match it with the reference ENF database from the power grid to estimate the time the dominant ENF signal embedded in the recording. We then subtract the power signal corresponding to the time of recording of the dominant ENF signal from the audio recording. As the magnitude of power measurements and the actual embedding in the audio may differ, the subtraction is performed by estimating the appropriate scaling factor of the magnitude that makes the ENF signal of resulting audio signal $\hat{s}(t)$ maximally decorrelated with the ENF signal of the power recording corresponding to the time of dominant ENF embedding estimated previously. More specifically, we have:

$$\begin{aligned} \hat{s}(t) &= s(t) - \hat{a} \cdot P(t), \quad with \\ \hat{a} &= \underset{a}{argmin} \left\{ corr(ENF(s(t) - aP(t)), ENF(P(t))) \right\}, \end{aligned} \tag{4.2}$$

where $P(t)$ is the power measurement signal at time t, and $\hat{a}$ is the estimated magnitude of the power. $ENF(\cdot)$ denotes the weighted frequency estimation function. As can be understood from the equation, the selection of $\hat{a}$ is to search for the relative amplitude of the dominant ENF signal in the audio signal, with respect to the power signal. Ideally, after the decorrelation process, the resulting signal $\hat{s}(t)$ is contains no traces of $E_d(t)$. The ENF signal that remains in $\hat{s}(t)$ would come from the latent ENF signal, $E_l(t)$. We estimate the latent ENF using the weighted frequency estimation approach from $\hat{s}(t)$.

To show the effectiveness of the proposed algorithm in extracting the dominant

and the latent ENFs, we conduct experiments on audio data. An audio recording was made in an acoustic anechoic chamber and recaptured later in the same place. The power measurements were also recorded during the original recording and the recapturing process. The original ENF and the recapturing ENF signals are present around 120 Hz in this case. The ENF extracted directly from the recaptured audio signal shows similar fluctuations with the ENF signal estimated from the power signal at the time of the original recording (NCC 0.62), as can be seen from Fig. 4.4. The content ENF signal is, therefore, the dominant signal in this case. We then decorrelate the recaptured audio by subtracting the estimated dominant ENF signal, as discussed previously. The ENF signal in the power measurement recording is centered around 60 Hz, so we transfer it to 120 Hz by squaring the power signal and feeding it into a bandpass filter with a narrow passband around 120 Hz. The processed power signal is used for decorrelation as in (3). The ENF signal estimated from the decorrelated audio signal shows high correlation with the ENF signal extracted from the power measurements at the time of recapturing (NCC 0.68). Both the original ENF and recapturing ENF are now successfully extracted from the recaptured audio recording using the proposed decorrelation method. The time of the original recording and recapturing can then be determined by comparing the original ENF and recapturing ENF with the reference database obtained from the power mains.

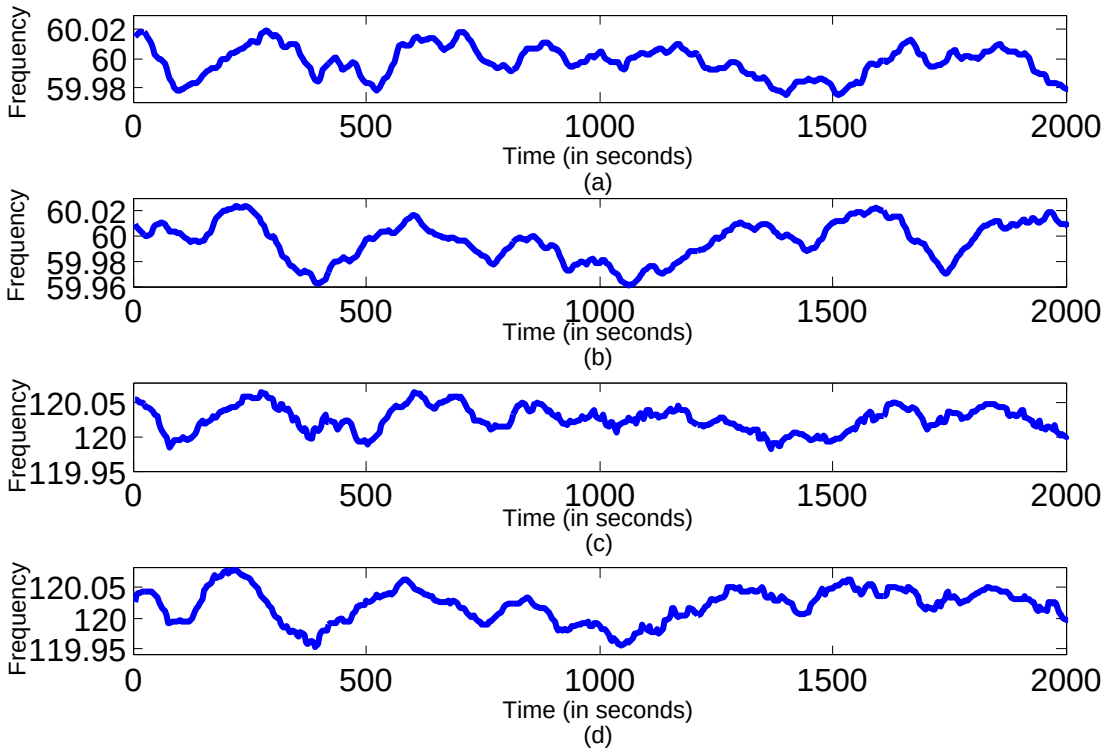


Figure 4.4: ENF fluctuations as a function of time. (a) the ENF from the power measurement signal at the time of the original recording; (b) the ENF from the power measurement signal at the time of recapturing; (c) the dominant ENF estimated from the recaptured audio recording; (d) the ENF estimated from the decorrelated audio signal.

Applications for Recapture Detection

Assuming that the reference measurements from the power grids are available, the proposed decorrelation algorithm can be used for audio recapture detection, i.e., identify whether the given audio recording is original or a recaptured version. As discussed earlier, two ENF signals are embedded in a recaptured audio recording. The ENF signal estimated directly from a recaptured audio recording is the dominant ENF signal E_d. After decorrelation, the latent ENF signal E_l can be extracted

from the residue audio signal. When compared with the reference ENF database measured from power mains, these two ENF signals should match different segments of the reference database, the time indices of which are denoted as T_d and T_l, respectively.

If the recording is original, T_d and T_l are likely to be of similar values. In cases where T_d and T_l differ greatly, the peak correlation C, that is calculated between the ENF signal estimation from the decorrelated audio signal and the reference database, should be low since it is a false match. Under a hypothesis framework, the H_1 and H_0 cases and the decision rule can be formulated as follows:

$$\begin{cases} H_1: & \text{Test audio is recaptured.} \\ H_0: & \text{Test audio is original.} \end{cases}$$

$$\mathbb{1}(|T_d - T_l| > \delta) \times C \underset{H_0}{\overset{H_1}{\gtrless}} \tau$$

Here $\mathbb{1}(\cdot)$ is an indicator function, and τ is a decision threshold.

We conduct the following experiments to evaluate the proposed audio recapture detection scheme. Audio recordings were made in the acoustic chamber and a conference room. Some recordings were then recaptured in the acoustic chamber by playing on a speaker with variant volume. The total test dataset includes 8.5 hours of original recordings and 16 hours of recaptured ones. The recordings are divided into short clips of 10, 20, and 30 minutes duration, and each clip is considered a test sample. We evaluate the false alarm rate and detection rate with different values of τ to obtain the ROC curves, as shown in Fig 4.5. The detection accuracy increases with longer clips. Specifically, when considering audio clips of 30 minutes, 95% of

the recaptured clips are correctly identified without any false alarms.

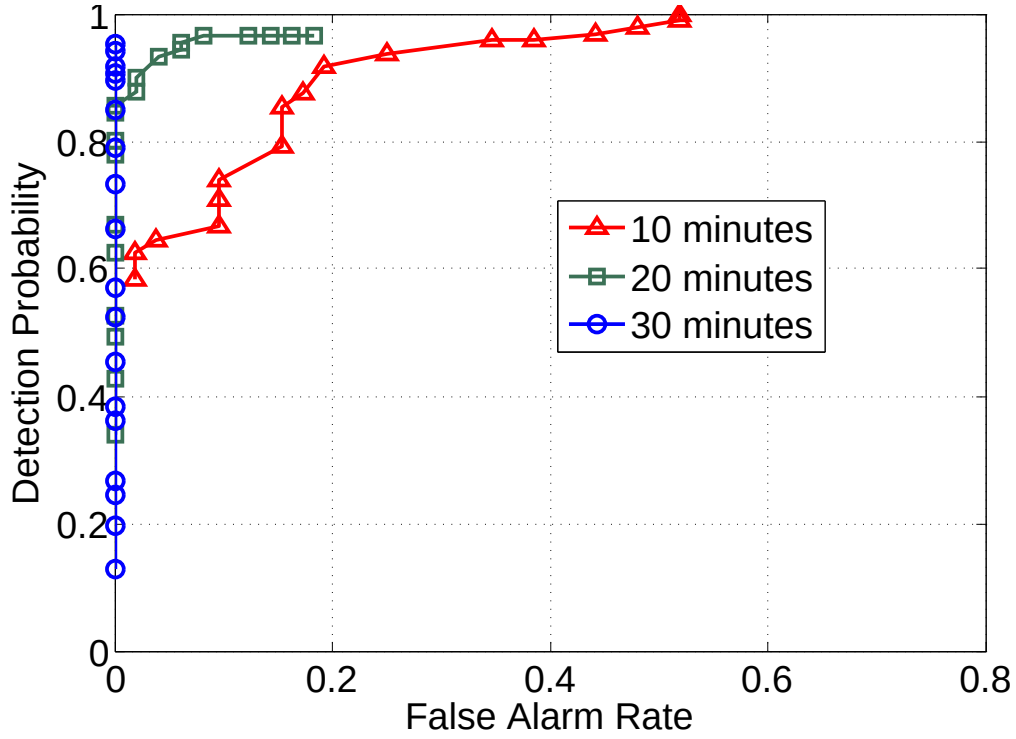
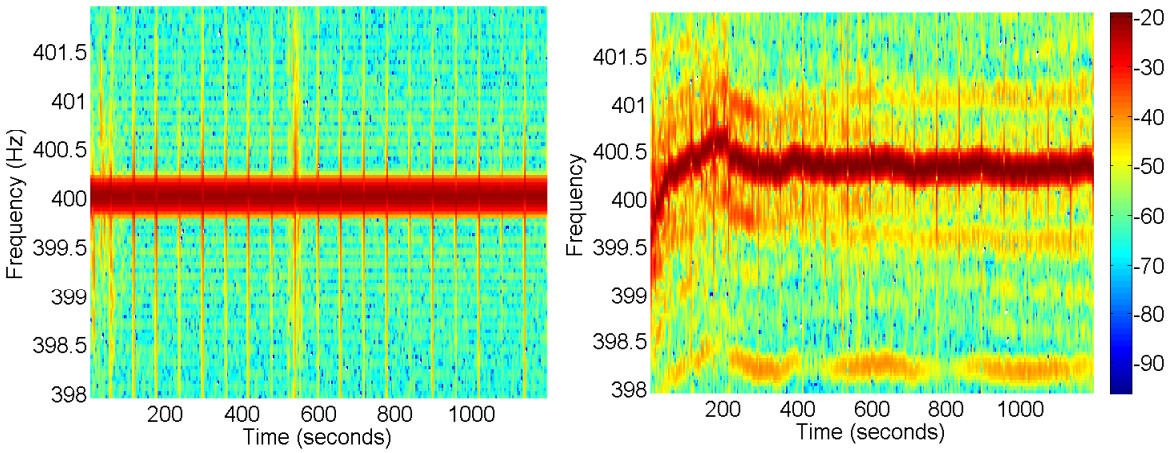


Figure 4.5: ROC for audio recapture detection with different clip lengths.

4.2.2 The Drifting Effect

Another challenge of dealing with digitized tape recordings is the "drifting" effect. This effect refers to the phenomenon that the signal frequency in a digitized analog audio recording may deviate from its original value. To demonstrate this, we conducted the following experiment with a cassette tape recorder. We generated a synthetic single tone acoustic signal of 400 Hz. We then made recordings of it with the tape recorder and a digital recorder, respectively. The tape recording was digitized by playing on a speaker and recording with a digital recorder in a acoustic chamber. The spectrograms of the recordings are shown in Fig. 4.6. As we

observe from the spectrograms, the digital recording captures the signal at almost exactly 400 Hz, while the signal frequency of the digitized tape recording presents random deviations from its original value. This effect occurs probably because that the mechanical rolling speed of the tape recorder is not identical during recording and playback. The signal presents at a higher frequency if the rolling speed during playback is faster than that during recording, and vice versa.



(a) (b)

Figure 4.6: Demonstration of the drifting effect with a synthetic 400 Hz tone signal. (a) Spectrogram of the digital recording; (b) Spectrogram of the digitized cassette tape recording.

As the ENF signal has a relatively small dynamic range (less than 0.1 Hz around 60 Hz), deviations caused by the drifting effect may have a large impact on the estimation of ENF signals. In Fig. 4.7, the blue curve shows the ENF signals estimated directly from the digitized tape recordings in our experiment. The ENF

signals estimated from the simultaneous power measurements (shifted to the same scale) is shown as the red curve in the figure. Due to the drifting effect, the audio ENF signals cannot match well with the groundtruth power ENF signals.

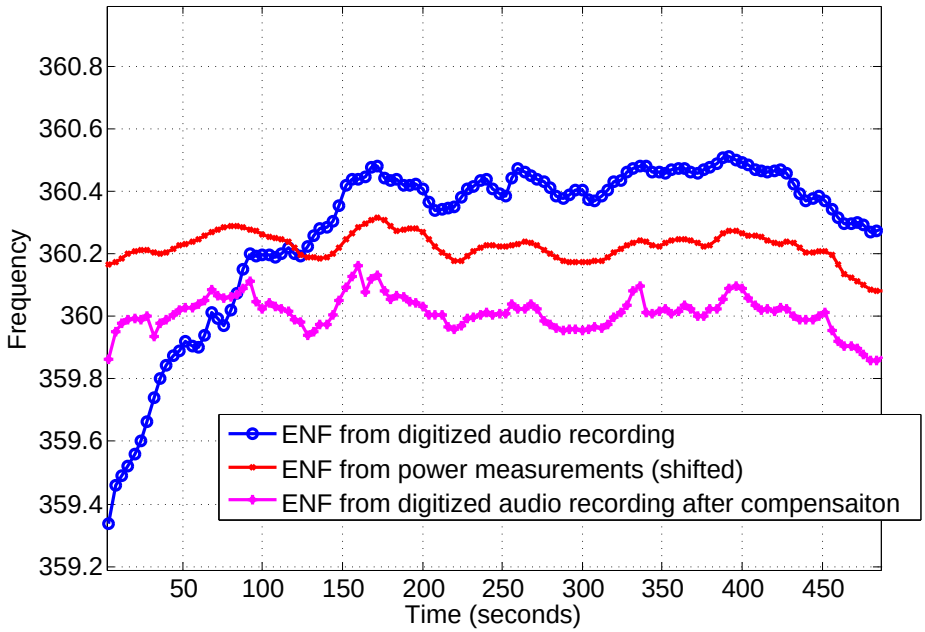


Figure 4.7: Compensating for the drifting effect.

We can compensate for the drifting effect if a reference tone signal of frequency f^t is available, as in our experiment (400 Hz). ENF_d denotes the ENF signal estimated directly from the digitized audio, and f_d^t is the frequency of the tone signal present in the recording. The adjusted ENF signal can be expressed as

$$ENF_c = \frac{ENF_d * f^t}{f_d^t} \tag{4.3}$$

In Fig. 4.7, the pink curve is the ENF signal after compensation, and it shows similar trends as the groundtruth power ENF signal.

The drifting effect is commonly seen among digitized historical recordings, such as the tapes of the NASA Apollo mission. Fig. 4.8 shows the spectrogram of a digitized Apollo Mission 11 audio recording. In the spectrogram, we can see three strips, two of which exhibit the same variations. The other strip resides at approximately 120 Hz. Based on previous subsection, we infer that the former are the original ENF signals and the latter is the recapturing ENF signals from the digitization process. The original ENF signals that originate from the analog tape recording deviate from the nominal value of 120 Hz because of the drifting effect. The discontinuity in the original ENF signals is caused by pauses during tape recording or deletion of certain content during digitization.

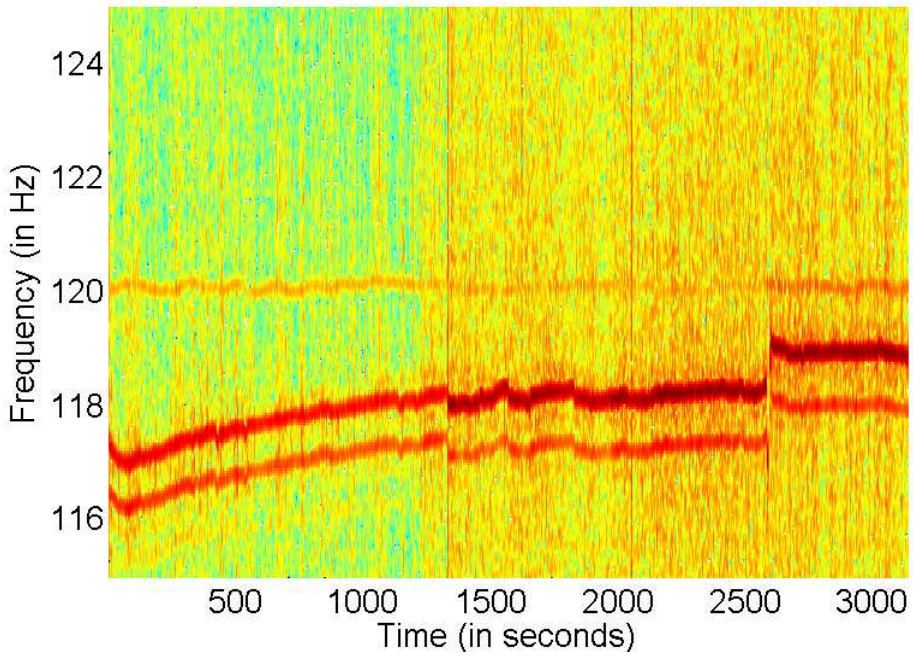


Figure 4.8: The spectrogram of an Apollo Mission 11 recording.

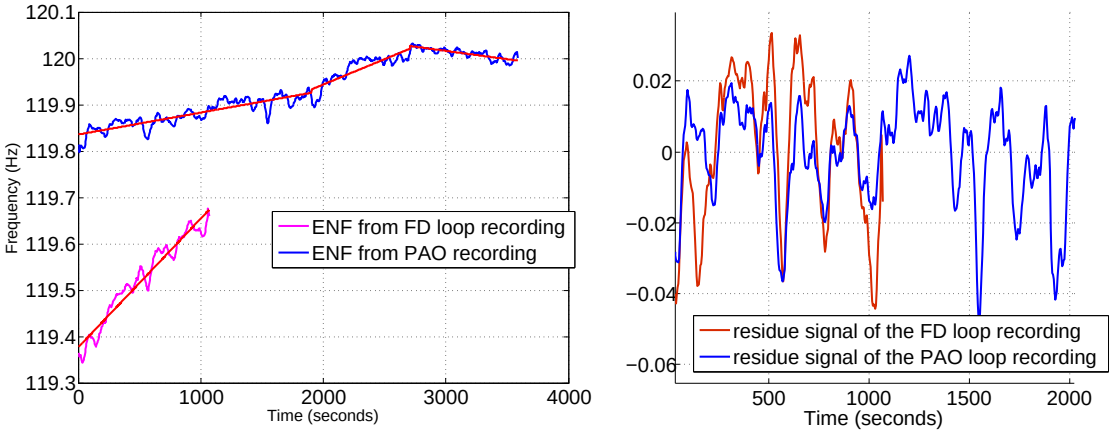
Different from the previously discussed scenario, the reference tone signal is

not available in those historical recordings. In some situations, the drifting effect can be approximated well with a linear or piece-wise linear model. We provide an example in Fig. 10. The ENF signals are extracted from two tapes in the Apollo 11 audio dataset [2] that were recorded at the same time. One of the recordings was made in the Flight Directors (FD) loop, and the other one was made in the Public Affair Officers (PAO) loop. These two recordings have been manually synchronized based on conversational content present in both recordings. Given the drifting effect, directly correlating the ENF signals from the two recordings fails to find a true match, which should be at the lag of 0 seconds. We fit the ENF signals with a piece-wise linear model, as illustrated in Fig. 4.9 (a). Fig. 4.9 (b) shows the residues between the original ENF signals and their linear approximations as new timing features, which exhibit consistent variation patterns over time. The correlation coefficients with different lags are plotted in Fig. 4.10. We can see that before the linear fitting and residue extraction, no correlation peak can be observed; and the correlation between the residue signals successfully finds the proper temporal alignment.

4.3 Audio Speed Restoration

4.3.1 ENF as a Guidance for Speed Correction

The mechanical rolling speed of an analog tape recorder and player is often not well stabilized. As discussed in Sec. 4.2.2, when an analog tape recording is digitized, the rolling speed of the tape playback likely to differ from that of the



(a) (b)

Figure 4.9: Using linear fitting to compensate the drifting effect. (a) The ENF signals from the two recordings and their liner approximation; (b) The residue signals.

creation of the tape. As a result, a speed offset is induced in the digitized version of the audio. Significant speed errors can sometimes affect the perceptual quality of the audio recording. Even with minor speed error that may not be audible to human ears, it is often desirable to perform speed correction. For example, many old recordings have great historical and archival value, so the digitized version of the recording should preserve the original audio as faithfully as possible. Certain historical events were covered by multiple audio recordings [1, 2], and it would be interesting to synchronize the recordings and play them simultaneously in multiple channels. Speed errors may hinder such an application, as it is difficult to align recordings that have unknown and different speed offsets.

As manual detection and correction of speed error in audio may be too costly and inefficient in practice, an automatic solution is preferred. We have designed a

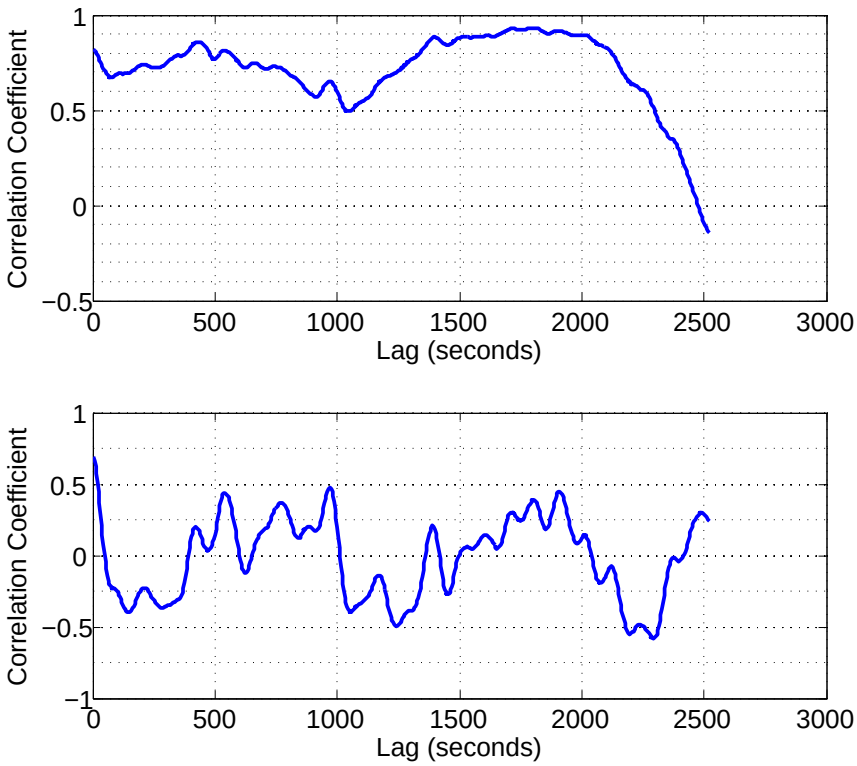


Figure 4.10: Correlation coefficients between the ENF signals (upper) and their residue signals (lower).

tape speed correction scheme exploiting ENF signals embedded in audio recordings.

As discussed in previous sections, the nominal value of the ENF is known (60Hz in North America, 50Hz in most other parts of the world). The instantaneous value of the ENF typically fluctuates around its nominal value as a result of the interaction between power load and generation, and the deviation of the ENF from its nominal value is usually minuscule given the control mechanism of the power grid. For example, in the US the ENF deviation is typically less than 0.05Hz from the nominal value of 60Hz, which is equivalent to about 0.08%. The ENF signal can

be approximately considered as a single tone signal. If an original analog recording has captured ENF traces during its creation, and such traces could be retained in the digitized version of the recording, then we can exploit the ENF signal to detect speed offsets. The ENF signal appearing at a frequency higher than its nominal value indicates the audio speed is faster then normal, and vice versa. Further, the ENF signal may serve as a reference signal to perform speed error correction.

A general diagram of the proposed scheme is plotted in Fig. 4.11. The given audio signal is first divided into frames of unit length, and the ENF signal f is extracted for every frame. We refer to each frame as an ENF frame hereafter. A speed correction ratio is obtained as

$$r = f/f_0, \tag{4.4}$$

where f_0 is the nominal value of ENF. Speed correction is then performed by temporally stretching or compressing of the audio signal.

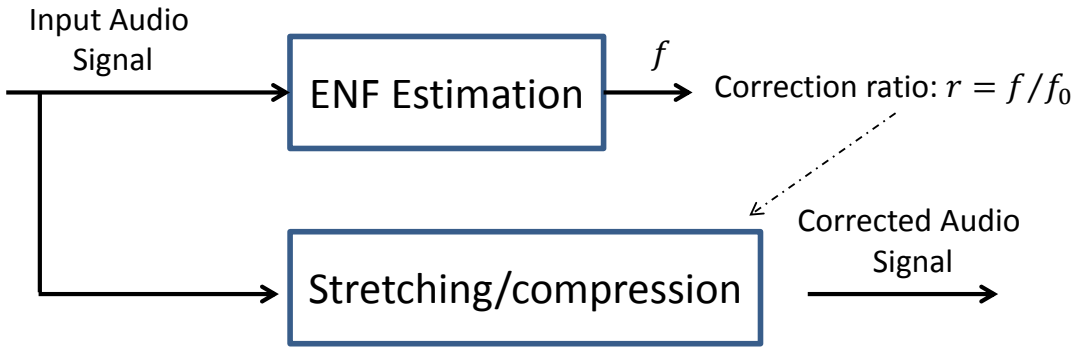


Figure 4.11: The diagram of the ENF-based speed correction scheme.

The values of ENF signal in adjacent ENF frames are likely to be different from each other. This would cause a sudden speed change in the restored audio at

the frame boundaries. To alleviate such artifact, each ENF frame is further divided into several smaller chunks, referred to as correction frames. The ENF signal for the correction frames are obtained via interpolation among the ENF signals of the ENF frames. Therefore, the values of the ENF signal transit smoothly from one ENF frame to the next, and each correction frame has its unique correction ratio that differs slightly from its neighbors.

When choosing ENF frame size, a trade-off exists between the frequency estimation accuracy and temporal resolution. The optimal amount of speed adjustment of a recording may change quickly over time. Ideally the ENF frame should be as small as possible to achieve satisfactory temporal resolution. However, accurate frequency estimation requires that the ENF frame be of a certain length. Dividing an ENF frame into several correction frames may provide an improved trade-off between the frequency estimation accuracy and temporal resolution.

Within each correction frame, sample rate conversion is performed on the audio signal so it becomes temporally stretched or compressed according to the correction ratio r. Two methods of sample rate conversion are considered in this work, and their performances will be analyzed and experimentally compared.

One of the most common strategy of rate conversion is up-sampling followed by down-sampling [45]. The diagram of this scheme is shown in Fig. 4.12. The correction ratio is converted to a rational number $r = \frac{U}{D}$, with U and D being integers. The audio signal first goes through an U-fold up-sampler which inserts $U - 1$ zeros between every adjacent samples, followed by a low-pass filter. Then a D-fold down-sampler retains 1 out of every D samples, and the output is the

speed corrected audio signal. The low-pass filter between the up-sampler and down-sampler replaces zeros in the up-sampled signal in interpolated values, and avoids aliasing in the following decimation.

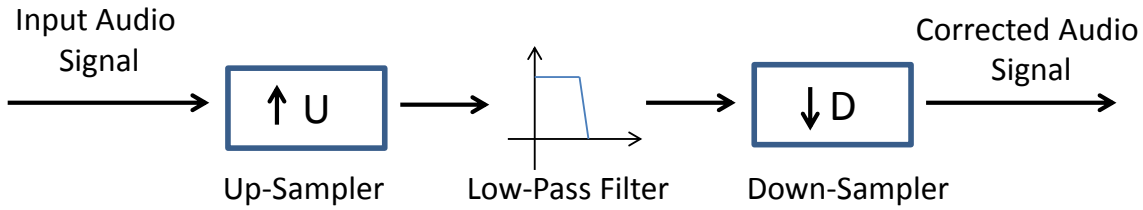


Figure 4.12: Speed correction based on up-sampling and down-sampling.

In practice, the correction ratio may not be a rational number. In other cases, U and D may be extremely large numbers resulting in extraordinary computational and storage complexity. Therefore, approximations are applied and the values of U and D are chosen as the smallest integers satisfying the following rule:

$$|\frac{U}{D} - r| < \delta \cdot r, \tag{4.5}$$

where δ is a pre-set parameter to adjust the trade-off between efficiency and precision.

Let f_0, f_t, f, f_r denote the nominal value of ENF, the true value of ENF, the observed value of ENF in the audio before correction, and the observed value of ENF in the restored audio, respectively. From the discussion above, we have

$$r = \frac{f}{f_0};$$

$$\hat{r} = r + \Delta r,$$

$$f_r = \frac{f}{r_q},$$

where Δr is the rational approximation error introduced from Eq. 4.5. Assume that the true value of ENF deviates from the nominal value by Δf:

$$f_t = f_0 + \Delta f.$$

The accuracy of the speed correction can be measured by

$$\begin{aligned} E &= 1 - \frac{f_r}{f_t} \\ &= 1 - \frac{f}{\hat{r} \cdot f_t} \\ &= 1 - \alpha. \end{aligned} \tag{4.6}$$

Here,

$$\begin{aligned} \alpha &= \frac{f}{\hat{r} \cdot f_t} \\ &= \frac{f_0 \cdot r}{(r + \Delta r)(f_0 + \Delta f)} \\ &= \frac{1}{(1 + \frac{\Delta r}{r})(1 + \frac{\Delta f}{f_0})} \end{aligned} \tag{4.7}$$

From Eq. 4.7, we can see that the accuracy of the speed correction is affected by two factors: the deviation of the ENF signal from its nominal value (Δf) and the approximation error in the correction ratio (Δr). The analysis above is based on the assumption of perfect ENF estimation. If this assumption does not hold, the speed adjustment is subject to additional distortion due to frequency estimation errors.

The other sample rate conversion method is based on sample interpolation, as has been used in [47]. The original audio signal is denoted by $s[n]$, $n = 1, 2, ...$

With a given correction ratio r, the ideal corrected audio signal is $s_{opt} = s[r \cdot n]$. Here, r may be an arbitrary positive real number, so the product of r and n is likely not an integer. For a non-integer x, the value of $s(x)$ can be obtained with nearby samples at integer indices using interpolation. In this work, we demonstrate this idea with a simple linear interpolation method. Assume the largest integer that is less or equal to x is $\lfloor x \rfloor = k$. With a linear interpolation as illustrated in Fig. 4.13, $s(x)$ is determined by

$$s(x) = w_1 \cdot s(k) + w_2 \cdot s(k+1),$$

where $w_1 = k + 1 - x$, and $w_2 = x - k$.

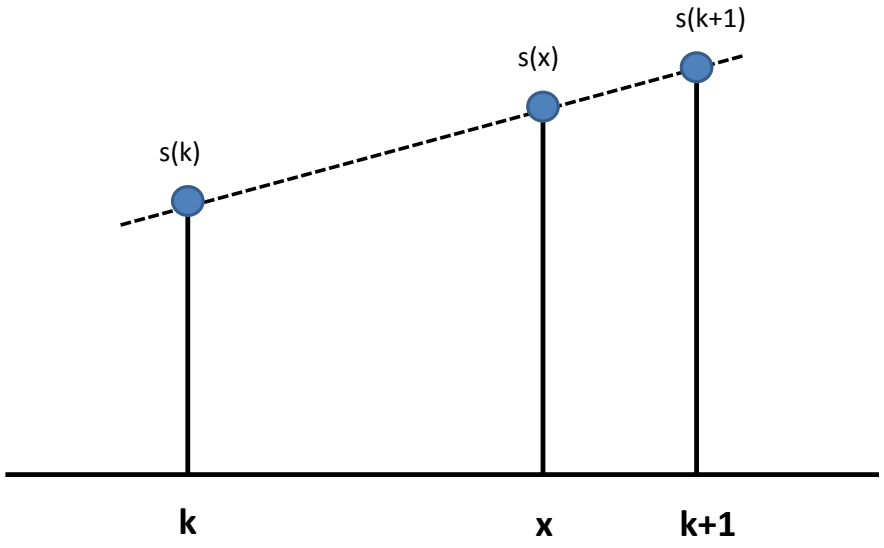


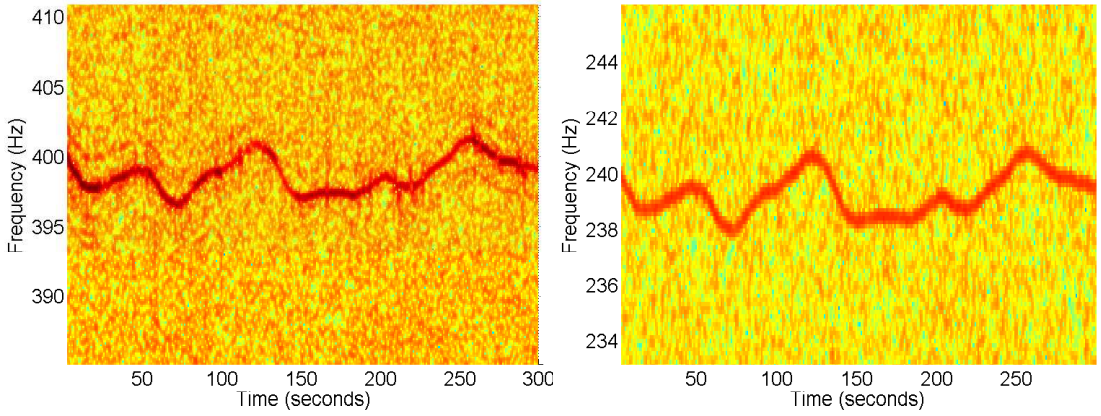
Figure 4.13: Linear interpolation for tape speed adjustment.

4.3.2 Experiments and Examples

Experiments have been conducted in order to evaluate the proposed audio speed correction approach and compare the two sample rate conversion methods. Audio recordings are created using a digital recorder in an office, and a synthetic tone signal of 400 Hz is generated during the recording. The speed of the recordings are then intentionally altered following random patterns via stretching and compressing of the audio signals. We extract ENF signals from the altered audio recordings and apply the proposed speed correction approach. Frequency analysis is performed on the audio signals before and after correction. The synthetic signal of 400 Hz provides a way of measuring the speed correction accuracy. If the audio speed is perfectly restored, the tone signal should appear at exactly 400 Hz.

A total of 100 tape recordings are used in our experiment, each being about five minutes long. The proposed speed correction approach is applied to all the audio signals using both rate conversion methods. The size of the ENF frame and correction frame are set as four seconds and one second, respectively. The ENF signals for the correction frames are obtained by linear interpolation of the ENF estimates of the ENF frame.

Fig.4.14 shows the spectrogram of an audio signal before speed restoration. Both the tone signal and the ENF signal have been distorted by the tintometric tampering with the audio speed. Fig.4.15 shows the spectrogram of the audio signal after speed restoration. Our approach is effective, and the tone signal is resorted close to its ground truth value.



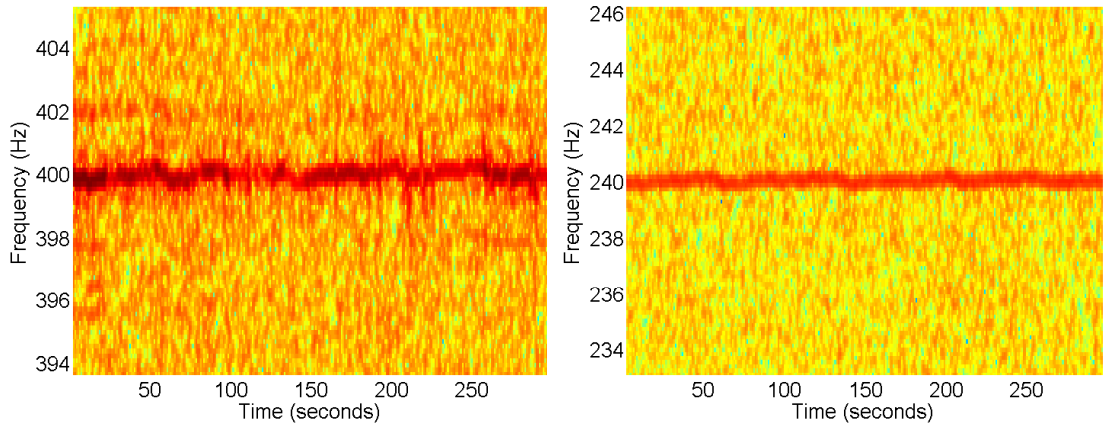
(a) (b)

Figure 4.14: The spectrogram of a test audio before correction. (a) The 400 Hz tone signal. (b) The ENF signal.

The effectiveness of the speed correction is measured by the closeness of the observed tone signal to its groundtruth value of 400 Hz. The restored audio signal is divided into frames of four seconds, and the dominant frequency around 400 Hz in the i^{th} frame is estimated as $tone(i)$, $i = 1, 2, ..., N$. The overall speed correction error is defined as root-mean-square-error (RMSE) of $tone(\cdot)$:

$$e = \frac{1}{N} \sum_{i=1}^{N} \sqrt{(tone(i) - 400)^2}. \tag{4.8}$$

Two rate conversion methods are tested in the experiment: re-sampling with up-sampler and down-sampler, and interpolation. For the re-sampling method, the error correction ratio is converted to a rational number, according to Eq. 4.7, in order to avoid excessive computational and memory cost. The average values of RMSE of the tone signal over all test audio clips are calculated and listed in Table 4.1, along with the average time needed for restoring a test audio clip with our implementation.



(a) (b)

Figure 4.15: The spectrogram of a test audio after correction. (a) the 400 Hz tone signal. (b) the ENF signal.

The experimental results demonstrate that the accuracy of speed restoration with the re-sampling method is improved using smaller approximation threshold δ, at the price of more the computational cost. The interpolation method for rate conversion generally outperforms the re-sampling method in both accuracy and efficiency, and therefore is chosen as the default method in the experiments in this chapter.

We have also conducted experiments to test the speed correction accuracy under different settings of values of ENF frame size L_e and correction frame size L_c. The results are listed in Table 4.2, from which several observations can be made. The speed offset of the audio may change quickly over time. Within each correction frame, a constant correction ratio is used to stretch or compress the audio signal. It is therefore desirable to adopt small correction frames to improve the speed correction accuracy. As for the size of ENF frame, it should not be either too large

Table 4.1: Comparison of the speed correction error using different rate conversion methods.

Rate conversion method	average RMSE	average time (sec.)
re-sampling with $\delta = 0.01$	1.81	1.04
re-sampling with $\delta = 0.005$	0.83	1.74
re-sampling with $\delta = 0.001$	0.15	1039
interpolation	0.12	0.103

or too small. A single dominant frequency value is estimated for each ENF frame, so the ENF frames should be reasonably short to adapt to the evolvement of the ENF signal over time. however, the number of samples in each ENF frame needs to be sufficient for reliable frequency estimation.

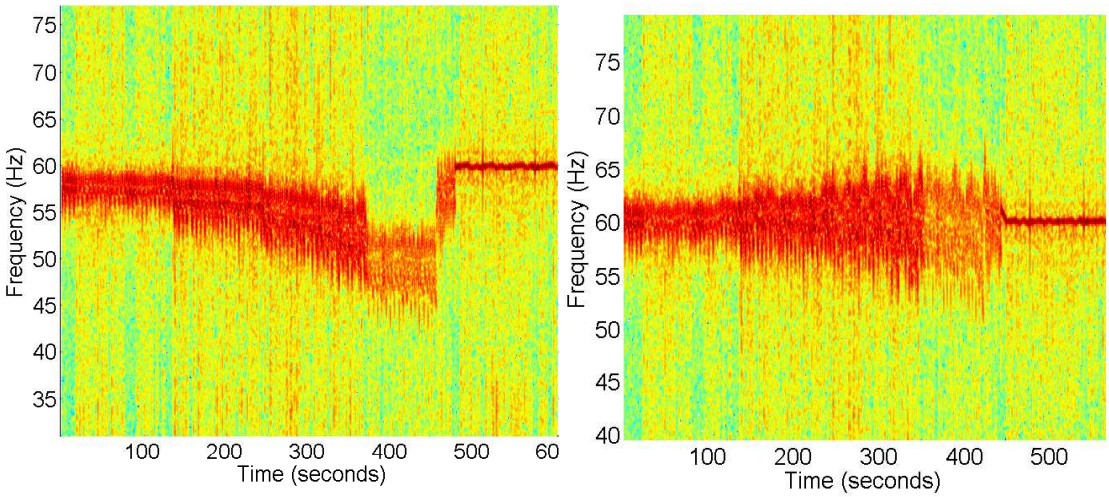
At the end of this section, we demonstrate the application of the proposed speed correction approach for restoration of certain digitized Apollo 11 recordings. The Apollo 11 mission, operated by NASA, had the objective to perform a crewed lunar landing and return to Earth. The mission launched on July 16, 1969, and finished on July 24, 1969. On July 20, the mission completed the first-ever human lunar landing. During the mission, a large amount of audio recordings were recorded of the communications among the crew of the spacecraft and the mission control staff. These recordings, made as historical documentation during the mission, were recorded on analog equipment that is now obsolete. When digitizing the recordings,

Table 4.2: Comparison of the speed correction error using different sizes of ENF frame (L_e) and correction frame (L_c).

$L_e \backslash L_c$ (sec.)	1	2	4	6	8
1	0.66				
2	0.17	0.19			
4	0.12	0.14	0.21		
6	0.16	0.18	0.23	0.30	
8	0.22	0.23	0.27	0.33	0.40

these machines no longer operate at the correct speed even under optimum circumstances. In some extreme cases, the tape is so greatly off speed that the recorded speech becomes unintelligible. Speed correction is therefore needed to preserve these tapes of tremendous archival value.

The spectrogram of a sample recording with serious speed offset is shown in Fig. 4.16 (a). From the figure, we observe the capture of ENF traces around 60 Hz. The value of the ENF signal decreases gradually to as low as 45-50 Hz in the earlier part of the recording, and then returns to close to 60 Hz through to the end. The proposed speed restoration scheme is applied on this recording using the ENF signals estimated around 60 Hz that is shown in Fig. 4.17 (a). The spectrogram and the estimated ENF signal from the audio after restoration is shown in Fig. 4.16 (b) and Fig. 4.17 (b), respectively.

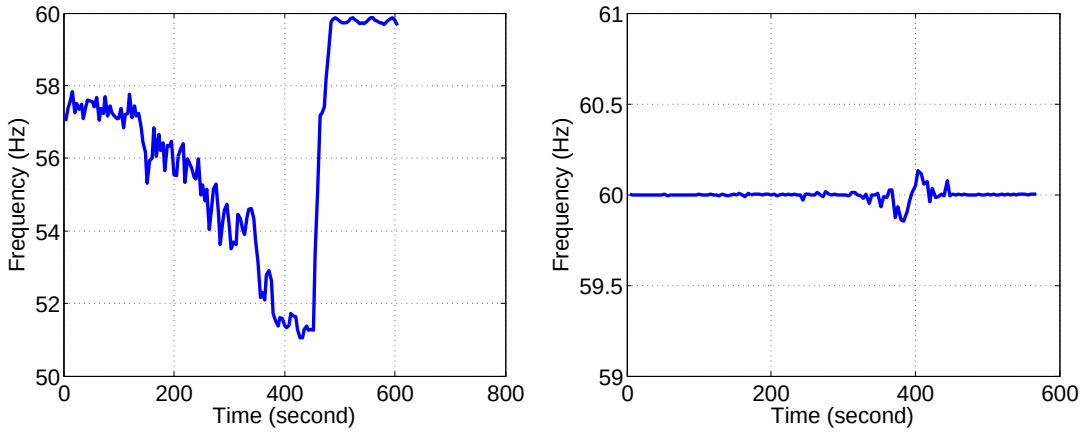


(a) (b)

Figure 4.16: The spectrogram of an Apollo mission control recording before and after speed correction. (a) Before speed correction; (b) After speed correction.

One way of measuring the speed correction accuracy with the Apollo recordings is to examine the Quindar tones in the audio. The Quindar tones have been used in Apollo Missions for Mission Control to simulate the action of the push-to-talk (PTT) and release-to-listen button. Two tones exist as pure sine waves that were 250 ms long but at slightly different frequencies around 2500 Hz. The "intro tone" is generated at 2525 Hz and signals the keypress of the PTT button that un-mutes the audio. The "outro tone" is generated at 2475 Hz and signals the release of the PTT button that mutes the audio.

Most Quindar tones sound loud and clear in the recording, resulting in a high signal to noise ratio. Their frequency therefore can be accurately estimated. The



(a) (b)

Figure 4.17: The ENF signal extracted from a Apollo mission control recording, (a) before and (b) after speed correction.

amount of deviation of the Quindar tones from their nominal values indicates the level of speed offset in the audio. For the audio clip whose spectrogram is shown in Fig 4.16 (a), Quindar tones' frequencies before and after speed correction is shown in Fig. 4.18. The frequencies of the Quindar tones in the original audio during the first 400 seconds are generally much lower than the nominal values of 2525 and 2475 Hz. The worst one is at about 2050 Hz, which is more than 17% lower than the nominal value. A measurement of the Quindar tone deviation can be defined in the following form:

$$d_Q = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_i - 2500)^2},$$

where Q_i is the frequency of the i^th of a total of n Quindar tones. The Quindar tone deviation in the original audio is 176. After restoration, the offset of the

Quindar tone frequencies is reduced, as can be seen from Fig. 4.18, and the Quindar tone deviation decreases to 60.8. This demonstrates that our proposed approach is effective in reducing the speed offset.

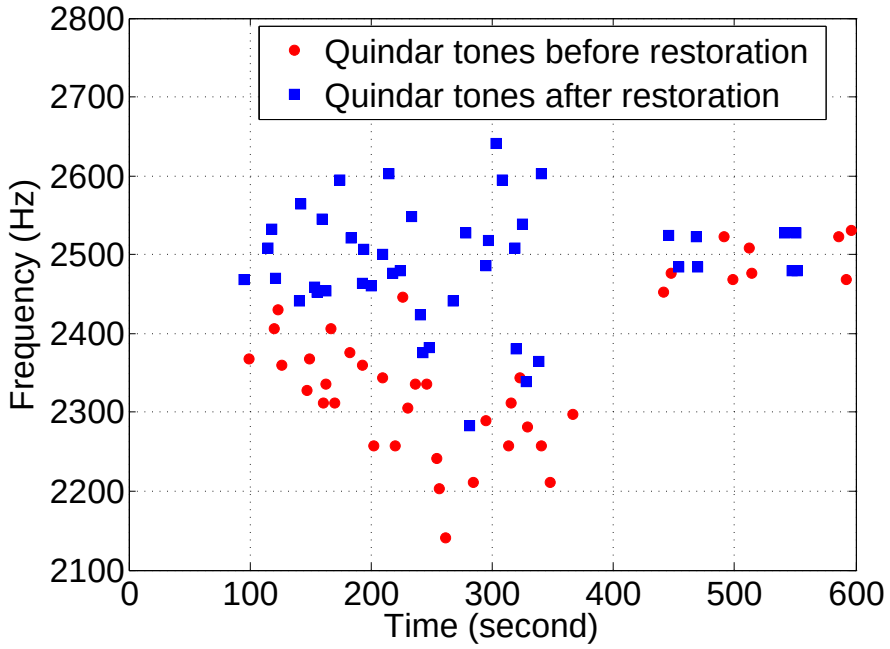


Figure 4.18: The Quindar tones in an Apollo Mission recording before and after restoration.

As indicated by the Quindar tone deviation, although the speed offset is reduced, the audio speed is not perfectly recovered. This is because that the ENF signal cannot be well estimated. Two factors contribute to this issue. First, from the spectrogram we observe that the ENF traces in the recording spread to a wide frequency strip, hindering accurate estimation of the dominant frequency. Second, it can be inferred from the Quindar tones in Fig. 4.18 that the speed offset of this recording changes quickly with time: some of the Quindar tones that occur close in

time have very different frequencies. As a result, the value of the ENF signal in the recording also changes quickly. As the frequency estimation must be performed on an ENF frame of sufficient length, the temporal resolution of the ENF estimation is not sufficient to cope with such fast changes.

4.4 Chapter Summary

In this chapter, we studied the ENF analysis on recaptured audio recordings. As ENF signal is embedded into multimedia recordings at the time of recording, multiple ENF traces may exist in a recording that has been recaptured. If the recapturing of the recording is performed in the region of the same nominal ENF as the original recording, the ENF traces of the two recording processes may overlap. Conventional frequency estimation methods may fail in such situations. We propose a decorrelation based algorithm to estimate multiple ENF signals from a recaptured audio in a sequential order, assuming that the power measurements are available as reference.

ENF extraction from recaptured recordings are relevant to analyzing digitized old recordings of historical importance. Many audio recordings that record historical events and conversations were created using analog recorders. These recordings are often digitized to facilitate preservation and transmission of the audio. During the digitization, a new ENF signal may be captured. A particular distortion for the ENF signal in digitized audio recordings is the drifting effect. It refers to the phenomenon that the value of the ENF in a digitized analog audio recording may deviate from

its original value. The mechanical rolling speed of analog tape recorder and player is often not well stabilized. When an analog tape recording is digitized, the rolling speed of the tape is likely to differ from that of the tape's creation. As a result, a speed offset is induced in the digitized version of the audio, causing the drifting effect.

We propose to correct speed error of digitized audio recording using ENF as a reference signal. The ENF signal fluctuates minimally around a known nominal value. It therefore can be considered as a single tone signal. The ENF signal appearing at a frequency higher than its nominal value indicates the audio speed is faster then normal, and vice versa. Speed error correction can be performed by stretching or compressing the audio signal according to the amount of deviation of the ENF signal. The proposed method has been applied to audio recordings from the NASA Apollo 11 Mission, and it has been shown to be effective.

Acknowledgement

I would like to thank Prof. Douglas W. Oard of the University of Maryland for enlightening discussions that inspired the work in this chapter.

Chapter 5

Quality Evaluation of Image Features for Efficient and Robust Spatial Alignment

In the previous chapters, we have discussed audio/video synchronization that presents a fundamental task for applications involving multiple pieces of audio-visual data, such as video panorama, 3D reconstruction, and video superresolution. Spatial alignment of images and video frames, which is often performed together with or following synchronization, is another essential alignment problem for these applications. Spatial alignment is also useful for identifying common objects in images, facilitating such applications as image retrieval. Local image feature offers one of the most successful and popular methods for visual matching of images. In this chapter, we investigate feature quality evaluation for efficient and robust spatial alignment of images/video frames.

5.1 Chapter Introduction

5.1.1 Motivation

With the rapid growth of storage capacity and the development of manufacturing technologies, digital cameras are becoming ubiquitous, and the quantity of digital images is increasing drastically. The existence of common objects in these images have motivated research in content-based image retrieval (CBIR). Given a query image, the goal is to retrieve from a large database the images containing the same object or scene as in the query image. CBIR can be useful in scenarios including landmark recognition, direction aid for tourists, and CD/book automatic annotation. A special application of CBIR is mobile visual search [26], in which users can use their phone cameras to take a picture of an object, such as landmarks, and send the picture or the extracted image features via broadband data connections to a server. The server then conducts image retrieval and object recognition to identify the object in the picture and send desired information about the object, such as annotations, back to the user. Examples of commercial products in this category include Google Goggles [3], Nokia Point and Find [4], and Amazon SnapTell [5]. Mobile visual search poses unique challenges. The retrieval process must be conducted within stringent memory, computation, power, and bandwidth constraints. The communication and processing latency should be minimized to ensure satisfactory user experience. This means that the size of data transmitted between the mobile client and the server should be as small as possible, and the

image retrieval should be as computationally efficient as possible.

Most of state-of-the-art solutions for CBIR are built on the usage of local image features. In the local feature framework, interest points are first selected as distinctive and robust points in the image by a key point detector. Next, a robust feature descriptor is generated using the information within the neighborhood of the interest point. The images may have gone through different lighting and viewing conditions, so desirable features should be robust to various distortions. Among all the proposed local image features [40, 52], the Scale-Invariant Feature Transform (SIFT) [38] is regarded as a seminal work given its robustness. The interest points used in SIFT are extrema in the Difference-of-Gaussian scale space, and the descriptors are vectors consisting of the histograms of the weighted gradient orientations in a local patch around every interest point. The histogram is calculated on a 4-by-4 block for 16 such blocks in the neighborhood of the interest point. Each histogram has 8 bins covering 360 degrees, giving a SIFT descriptor of 128-dimension. The patches are orientation and scale normalized, so SIFT is resilient to rotation and scaling variations.

A query image can be matched against an image database by computing the pair-wise distances between the SIFT descriptors. However, this scheme is computationally expensive for large corpora and does not scale well. To cope with large-scale databases, a visual dictionary-based Bog-of-Words (BoW) approach has been proposed by Sivic and Zisserman [55]. The idea is to quantize image features into a set of visual words as a codebook by using k-means clustering on many training features. A given feature can be mapped to its nearest visual word among the code-

book. The images are then represented by the frequency vector of the visual word occurrences. The similarity between two images is usually measured using the L_1 or L_2 distance between their visual word frequency vectors. During a query, the similarity score can be computed efficiently by using an inverted file system associated with the database. To improve the retrieval accuracy, Nister and Stewenius use the hierarchical k-means clustering to generate a vocabulary tree for a much larger BoW codebook [43]. Alternatively, Philbin, et al., propose approximate k-means utilizing randomized k-d trees to achieve a better partition of the descriptor space [48]. Other extensions of this approach include query expansion [18], hamming embedding [34], and soft quantization [49].

The visual dictionary-based approach has proven to be significantly more efficient than contentional methods. Much of the computational gain comes from the inverted file system that decreases the fraction of images in the database that need to be considered for computing the similarity scores. In order to expedite the retrieval process, the visual world frequency vector of each image should be as sparse as possible so that the number of image hits is minimized. To achieve efficient and accurate retrieval, it is desirable to select from among all available features a subset of the most reliable and informative features to represent each image. Feature selection is especially beneficial to mobile visual search for which communication and processing latency should be minimized to ensure satisfactory user experience. As bandwidth is a limited resource in wireless communications, smaller number of features help reduce the data transmission load and communication latency between the mobile client and the server. The computational complexity of image retrieval

is significantly reduced with fewer, yet more resilient, features. Not only does this reduce the server side processing latency, it also potentially enhances the search performance. Another possible benefit of feature selection is smaller feature storage cost which keeps the size of the feature database feasible for large-scale applications.

In this work, we focus on designing a scheme to evaluate the quality of SIFT features in terms of their robustness and discriminability. A quality score is assigned to every SIFT feature based on its contrast value, scale, and descriptor, using a quality metric kernel that is obtained in a one-time training phase. The quality score predicts how the usefulness of a SIFT feature is in describing certain object or scene. Feature selection can be performed by retaining features with high quality scores. We also show the proposed methodology can be generalized and applied to other local image features, such as SURF.

5.1.2 Related Work

Several techniques proposed in recent years have the potential to improve the efficiency and accuracy of feature matching for image alignment and retrieval. Turcot and Lowe [62] considered multi-view scenarios and proposed to select only a small subset of useful image features to reduce the number of visual descriptors and also improve the recognition accuracy. Their method is based on an unsupervised pre-processing step to retain descriptors that are geometrically consistent across multiple views. The descriptors are ranked according to the time of appearances in the training images, and the top ones are retained. This method was improved in

both speed and accuracy by Naikal, et al., [42] by using sparse Principal Component Analysis (PCA). Fritz, et al., [21] present an information theoretic feature selection framework by employing the selection criteria of $H(O|f) < \lambda$. Here λ is a threshold, and $H(O|f)$ is the estimated conditional entropy of the object O given the feature f, defined as $H(O|f) = -\sum_k P(O_k|f)log(P(O_k|f))$; $P(O|f)$ is the conditional distribution of the object class variable O given that feature f comes from O, and this distribution can be estimated from the feature matching results of the training images. Further, Baheti, et al., [10] associate each descriptor to a weight factor that represents the relative importance of the feature in the database. They devise a three step approach of feature pruning, including intra-object pruning to remove redundancies between the views of the same object, inter-object pruning to retain the most informative features across different objects, and keypoint clustering to further reduce the number of features. Feature descriptor compression [15, 16, 26] has also been proposed for compact storage and reduced network transmission. SIFT descriptor can be compressed to around 2 bits per dimension with almost no performance degradation.

The feature selection method presented in this paper works by assigning a quality score to every individual image feature and retaining the features with the highest scores. This scheme is not as database-specific as some of the existing methods. It can also be potentially combined with state-of-the-art techniques.

5.2 Quality Evaluation of SIFT Features

5.2.1 Definition of Quality Metric

Useful features should provide true matches more frequently than generate false matches. A natural way of measuring the quality of features is to analyze their behaviors statistically. In [57], we estimate the probability of true match and false match of SIFT features and use them as quality metric. We apply vector quantization to the 128-dimension SIFT descriptor space to generate a set of visual words. SIFT features are extracted from many training images and assigned to their nearest visual words. The training images then undergo certain synthetic transformations, such as rotation and blurring. Feature matching is conducted between every pair of original and transformed images. Since the transformation is synthetic, the ground truth of point correspondence is known. If n features exist in a certain visual word bin, and n_t of them are correctly matched while n_f of them are falsely matched, then the probabilities of true match and false match are estimated as $\frac{n_t}{n}$ and $\frac{n_f}{n}$. The quality score of the features f belonging to this visual word bin can be defined in the form:

$$Q(f) = R_{true}(f) - \alpha R_{false}(f), \tag{5.1}$$

where $R_{true}(f)$ and $R_{false}(f)$ are the true match rate and false match rate respectively, and α is a weighting parameter. Note the sum of $R_{true}(f)$ and $R_{false}(f)$ is not necessarily 1 because some features may not be matched to any other features at all. A simplified assumption for this method is that features with descriptors

quantized into the same visual word are of the same quality.

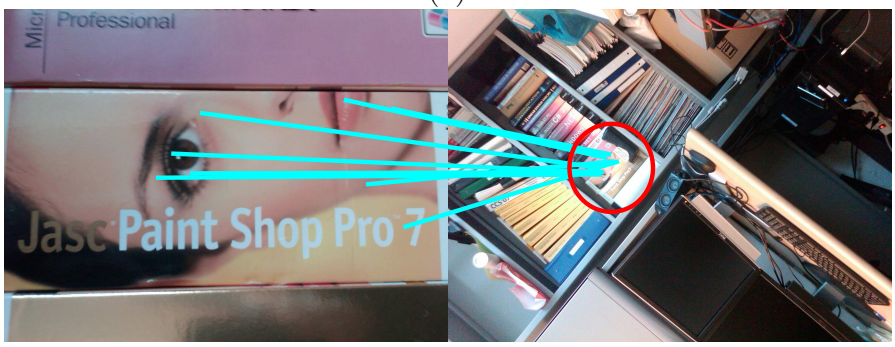
Fig. 5.1 shows an example of using the approach in [57] to prune SIFT features. The left figure shows the direct SIFT feature matching between the template image and an image containing the same object in the red circle, where many of the matches are incorrect. The right figure shows the matching result when using only those with good quality scores. Almost all false matches are eliminated while retaining all true matches.

Several heuristics to identify desirable SIFT features exist, such as selecting features with highest contrast value or largest scale [39,62]. The contrast value refers to the keypoint's response to the Difference-of-Gaussian filter at its characteristic scale. These empirical rules may also provide useful information for evaluating feature quality. In the following sections, we extend the work in [57] by incorporating contrast and scale information, and apply learning techniques to form a feature quality metric kernel. The whole contrast-scale-descriptor space is quantized into a set of bins, and each bin is characterized by its contrast quantization index, scale quantization index, and visual word index. We then analyze the statistical behavior of the features in each bin. We refer to the scale-contrast-descriptor space as the feature space.

5.2.2 Soft Quantization in Feature Space

The feature space quantization is necessary to make the feature quality evaluation feasible. Instead of assigning a quality score to every possible feature, we only

(a)



(b)

Figure 5.1: Comparing SIFT feature matching with and without feature pruning. The query object is shown on the left and the image containing the query object in the red circle is shown on the right. Figure (a) demonstrates matching result using all the extracted SIFT features; (b) demonstrates matching result using only selected features. The figures are best viewed in color.

need to deal with a limited number of feature space bins. However, the quantization provides only a relatively coarse approximation to the actual points in the feature space. The quality evaluation accuracy may be affected by quantization, especially for the features that fall near quantization boundaries. To alleviate the negative impact of quantization, we employ the soft quantization strategy, as in [49]. Instead of a hard quantization where a feature is assigned to a single bin in the feature

space, the soft quantization mechanism maps each feature to a set of weighted bins. The weights assigned to the bins can be determined by the distances between the feature point and the bin centroids. An exponential weighting is adopted in this paper.

Consider a SIFT feature f with scale S_f, contrast value C_f, and descriptor vector D_f. Suppose that $(S_{n_S}^{cen}, C_{n_C}^{cen}, D_{n_D}^{cen})$ is the feature space bin centroid corresponding to the n_S-th scale level, the n_C-th contrast level and the n_D-th visual word, where $n_S \in \{1, 2, ..., N_S\}$, $n_C \in \{1, 2, ..., N_C\}$, and $n_D \in \{1, 2, ..., N_D\}$, for a total of N_S, N_C and N_D quantization levels for scale, contrast, and descriptor, respectively. We denote $(S_{n_S}^{cen}, C_{n_C}^{cen}, D_{n_D}^{cen})$ by (n_S, n_C, n_D) in the following. The indices of the K_S nearest neighbors of S_f among $\{S_{n_S}^{cen} | n_S \in \{1, 2, ..., N_S\}\}$ are denoted as $\{I_t^S | t = 1, 2, ..., K_S\}$, i.e., the t-th nearest neighbor of S_f is $S_{I_t^S}^{cen}$. Similar notations are defined for the nearest neighbors of S_f and D_f (in L-2 distances). $\{I_t^C | t = 1, 2, ..., K_C\}$ and $\{I_t^D | t = 1, 2, ..., K_D\}$ are the nearest neighbor indices of S_f and D_f, respectively.

For hard quantization, f is assigned to a single feature space bin (I_1^S, I_1^C, I_1^D). With the soft quantization strategy, f is mapped to a set of $K_S \cdot K_C \cdot K_D$ nearby bins that are closest to (S_f, C_f, D_f), denoted by $\{I_{t_1}^S, I_{t_2}^C, I_{t_3}^D | t_1 = 1, ..., K_S, t_2 = 1, ..., K_C, t_3 = 1, ..., K_D\}$. We assign different weights to each bin according to the distance between the feature point and the bin centroids. Denoting the L2 distance

as $d(\cdot,\cdot)$, the weight assigned to bin $\{I^S_{t_1}, I^C_{t_2}, I^D_{t_3}\}$ is

$$\begin{aligned} w(I^S_{t_1}, I^C_{t_2}, I^D_{t_3}) &= \alpha_S \cdot \exp(-\frac{(d(S_f, S^{cen}_{I^S_{t_1}}))^2}{\sigma_S^2}) \\ &\quad \times \alpha_C \cdot \exp(-\frac{(d(C_f, C^{cen}_{I^C_{t_2}}))^2}{\sigma_C^2}) \\ &\quad \times \alpha_D \cdot \exp(-\frac{(d(D_f, D^{cen}_{I^D_{t_3}}))^2}{\sigma_D^2}) \\ &= w_S(I^S_{t_1}) \cdot w_C(I^C_{t_2}) \cdot w_D(I^D_{t_3}), \end{aligned} \tag{5.2}$$

where σ_S, σ_C and σ_D are tunable parameters. α_S, α_C and α_D are normalization coefficients defined as

$$\begin{aligned} \alpha_S &= \left[\sum_{t=1,2,\ldots,K_S} \exp(-\frac{(d(S_f, S^{cen}_{I^S_t}))^2}{\sigma_S^2})\right]^{-1}, \\ \alpha_C &= \left[\sum_{t=1,2,\ldots,K_C} \exp(-\frac{(d(C_f, C^{cen}_{I^C_t}))^2}{\sigma_C^2})\right]^{-1}, \\ \alpha_D &= \left[\sum_{t=1,2,\ldots,K_D} \exp(-\frac{(d(D_f, D^{cen}_{I^D_t}))^2}{\sigma_D^2})\right]^{-1}. \end{aligned} \tag{5.3}$$

The weights assigned to the bins that are not among the feature point's nearest bins can be thought of as 0. It can be shown that the sum of the weights across all the bins is unity:

$$\begin{aligned} &\sum_{t_1=1}^{N_S}\sum_{t_2=1}^{N_C}\sum_{t_3=1}^{N_D} w(t_1,t_2,t_3) \\ =\ &\sum_{t_1=1}^{N_S}\sum_{t_2=1}^{N_C}\sum_{t_3=1}^{N_D} w_S(t_1)w_C(t_2)w_D(t_3) \\ =\ &\left(\sum_{t_1=1}^{K_S} w_S(I^S_{t_1})\right)\left(\sum_{t_2=1}^{K_C} w_C(I^C_{t_2})\right)\left(\sum_{t_3=1}^{K_D} w_D(I^D_{t_3})\right) \\ =\ &1. \end{aligned} \tag{5.4}$$

5.2.3 Kernel Training and Quality Assessment

In this subsection, we explain how we integrate the techniques discussed above to form and utilize the SIFT feature quality metric kernel.

Quality Metric Kernel Training

To begin, we collect a set of representative sample images covering various scene types and categories as training images, and extract SIFT features from these training images. We then randomly select a subset of these features to estimate the quantization centroids for scale, contrast and descriptors, using K-means and Lloyd's algorithm. The numbers of quantization levels for each dimension, i.e., N_S, N_C and N_D, are determined empirically. The feature space is thus divided into bins that are indexed by their scale, contrast, and visual word indices. Then, we apply soft quantization to the features from the training images according to equation (2). The quantization weights are aggregated into a 3-dimension table N_{total}:

$$N_{total}(t_1, t_2, t_3) = \sum_{all\ \ features} w(t1, t2, t3), \tag{5.5}$$

where $t_1 \in \{1, 2, ..., N_S\}$, $t_2 \in \{1, 2, ..., N_C\}$ and $t_3 \in \{1, 2, ..., N_D\}$. We then conduct synthetic transformations of interest to the training images, and match SIFT features between every original and transformed image pair. The matching results for each feature from the original images are recorded. N_{true} and N_{false} are used to store the weight aggregates of each feature space bin corresponding to correctly and

falsely matched features, respectively:

$$N_{true}(t_1, t_2, t_3) = \sum_{correctly\ matched\ features} w(t_1, t_2, t_3),$$
$$N_{false}(t_1, t_2, t_3) = \sum_{falsely\ matched\ features} w(t_1, t_2, t_3). \tag{5.6}$$

The quality metric kernel can be established using the three tables above:

$$Q_{ker}(t_1, t_2, t_3) = \frac{N_{true}(t_1, t_2, t_3) - \alpha \cdot N_{false}(t_1, t_2, t_3)}{N_{total}(t_1, t_2, t_3)}, \tag{5.7}$$

where α is a parameter adjusting the weights between the true match rate and false match rate.

Quality Assessment

Given a SIFT feature f, we compute its quality score using the quality metric kernel. We first soft-quantize f to obtain the weights for each bin in the feature space according to equation (2), denoted by $w_f(t_1, t_2, t_3)$. Then its score is calculated as the dot product of $w_f(t_1, t_2, t_3)$ and the quality metric kernel Q_{ker}:

$$Q(f) = \sum_{t_1=1}^{N_S} \sum_{t_2=1}^{N_C} \sum_{t_3=1}^{N_D} w_f(t_1, t_2, t_3) \cdot Q_{ker}(t_1, t_2, t_3). \tag{5.8}$$

Features with higher quality scores are more desirable.

5.3 Experiment Results

5.3.1 Implementation and Experiment Setup

We collect about 40000 images from Flickr, covering various scene types and categories, as the training dataset. The resolution of these images is generally around

500-by-300. Some sample images are shown in Fig. 5.2. This dataset provides about 20 million SIFT features using the VLFeat SIFT implementation [65]. The number of quantization levels are chosen as $N_s = 10$, $N_c = 10$, $N_d = 4096$. 8000 images are randomly selected as the training images for estimating the quantization centroids and boundaries. We employ efficient hierarchical K-means to generate the visual word codebook. The transformations used for the quality metric kernel training and the corresponding parameters are listed in Table 5.1. The matching results between the original images and their transformed counterparts are used to learn the feature quality metric kernel as previously discussed. The parameters in Eq. 5.2 are set as $\sigma_S^2 = 2$, $\sigma_C^2 = 2$, $\sigma_D^2 = 10000$. The number of nearest neighbors for scale, contrast, and visual word are chosen as $K_S = 2$, $K_C = 2$, $K_D = 10$, respectively.

Figure 5.3 provides an example of feature selection on an image from the University of Kentucky Benchmark dataset [43]. The location of the features are indicated by the center of the red circles, and the radius of the circles are proportional to the feature scale. The first figure shows all the SIFT features extracted from the image. The other three figures demonstrate the features selected with highest quality score, largest scale, and highest contrast value, respectively. The example shows that the proposed quality score captures more informative and representative features than the other empirical schemes.

Table 5.1: List of transformations used in quality metric kernel Training

Transformation	Parameter	Value
Rotation	Rotation Angle	$15, 30, 45, ..., 165$
JPEG	Quality factor	80, 70, ..., 10
AWGN	Noise variance	0.01, 0.02, 0.03, 0.04
Average Filter	Filter size	3, 5, 7, 9
Gaussian Filter	Filter variance	0.5 1.0 , ... 3.0
Median filter	Filter size	3, 5, 7, 9
Histogram equalization	Output level numbers	256, 128, 64, 32

Figure 5.2: Sample images in the training dataset. The training dataset consists of around 40,000 images that are crawled from Flickr, including categories such as aircraft, cars, rocks, trees, etc.

5.3.2 Feature Selection Performance

To demonstrate the effectiveness of the proposed SIFT feature quality metric as a tool for selecting useful features, we conduct experiments on three datasets: the University of Kentucky Bechmark, the Oxford Buildings Dataset, and the INRIA Holidays Dataset. These are popular datasets in the literature and are often used as benchmarks for image retrieval. The three datasets are of different scales (10K, 5K, and 1.5K, respectively), and also provide a wide variety in terms of image content.

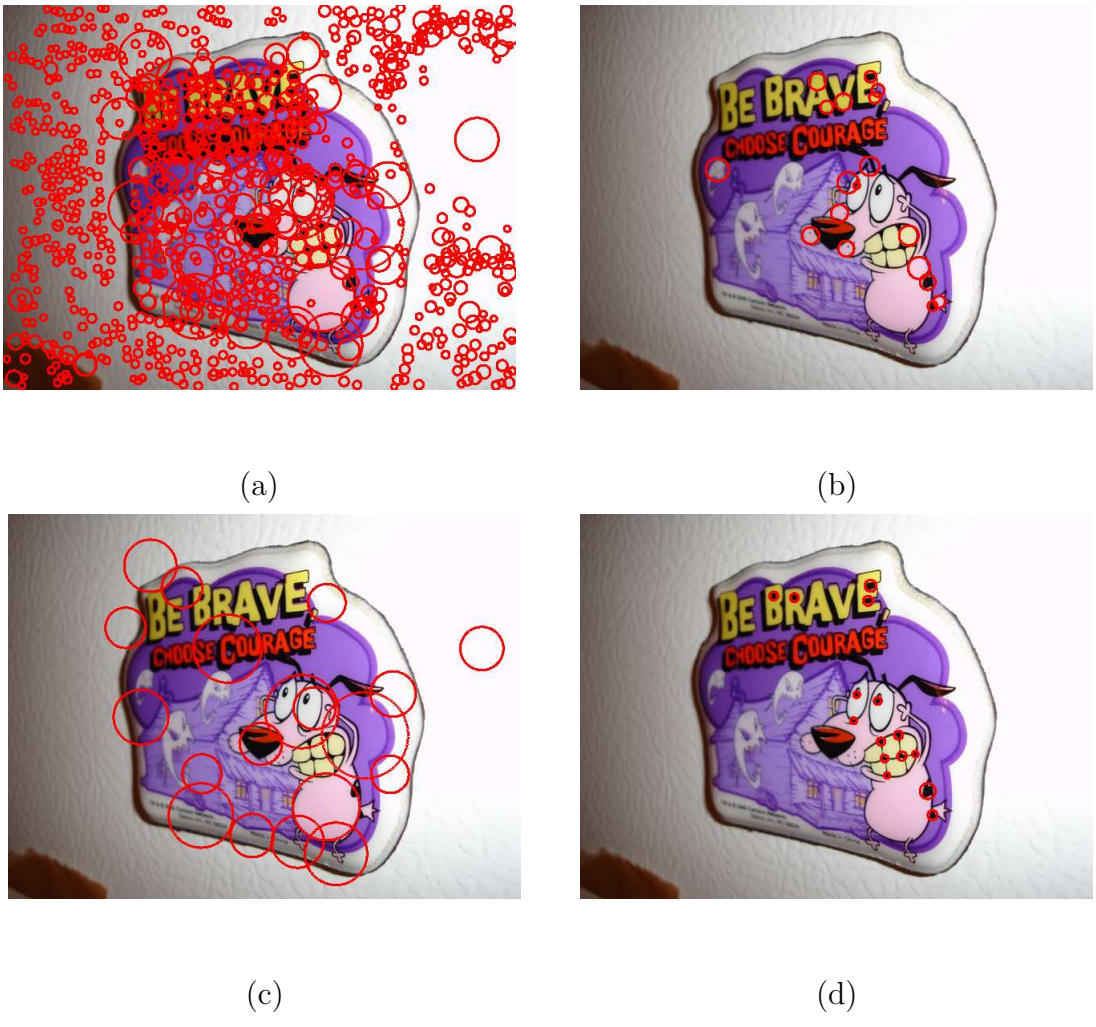


Figure 5.3: Comparing different feature selection schemes. (a) shows all the SIFT features extracted from the sample image; (b) shows the top 20 features with the highest quality scores; (c) shows the top 20 features of the largest scales; (d) shows the top 20 features with the highest contrast values. The features shown in red circles are best viewed in color.

The University of Kentucky Benchmark

The University of Kentucky Benchmark (UKB) [43] dataset consists of 10200 images of 2550 objects such as shoes, bottles, and CD covers. Each object is rep-

resented by 4 640×480 images from different viewpoints. The most popular performance metric for this database is the average top. Each image is employed as a query, and the perfect result provides all 4 images of the query object (including the query image itself) as top returned images. The average top is defined as the average number of correctly matched images that are in the top 4 returned images, taken over all possible query images.

We use the BoW approach in the experiments, and the vocabulary is obtained using the Hierarchical K-Means (HIK) method [43] with 1 million cluster centers. N SIFT features are selected to represent each image, and these features ($10200 \times N$) are assigned to their corresponding visual words. The similarity between images is measured with the tf-idf weighted scalar product between the frequency vectors of the visual word occurrences. An invert file system is built to facilitate efficient calculation of the similarity scores.

It is desirable to achieve a satisfactory average top score with the number of features N being as small as possible for efficient retrieval. Determining the best method of choosing the representative features for every image presents the key problem. With the proposed quality metric, we can calculate the quality scores of the SIFT features and choose the features with the highest scores. To show the effectiveness of our approach, we compare it with several other empirical feature selection schemes such as picking features with the largest scales or the highest contrast values. The average top scores under different settings of N are plotted in Figure 5.4.

In Fig. 5.4, the magenta curve with triangle symbols corresponds to the method

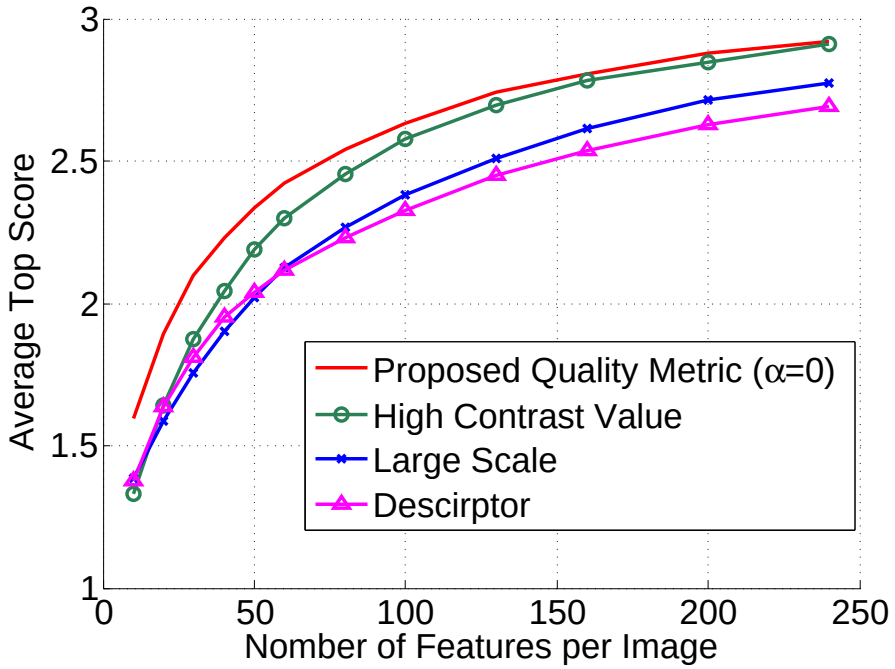


Figure 5.4: Average Top Scores on the UKB dataset using different feature selection schemes.

proposed in [57], where the quality metric is established considering only the descriptor information using hard quantization with 4096 visual words. The blue curve with cross symbols and the green curve with circle symbols show the performance of selecting by large scale and selecting by high contrast value, respectively. The red curve shows the performance using the proposed feature selection method (with $\alpha = 0$). The figure shows that the feature selection scheme, based on our proposed feature quality score, achieves superior performance when compared to other empirical methods, especially with smaller values of N, when the selection of useful feature is more crucial. Typically, a 20% saving of feature number can be obtained to achieve the same average top score. Detailed results are listed in Table 5.2.

We have also tested the effect of α in equation (7), as illustrated in Figure 5.5. The matching accuracy is not sensitive to the choice of α. This observation remains consistent over all three datasets we are testing on. In the following experiments, we consider the case when α is 0.

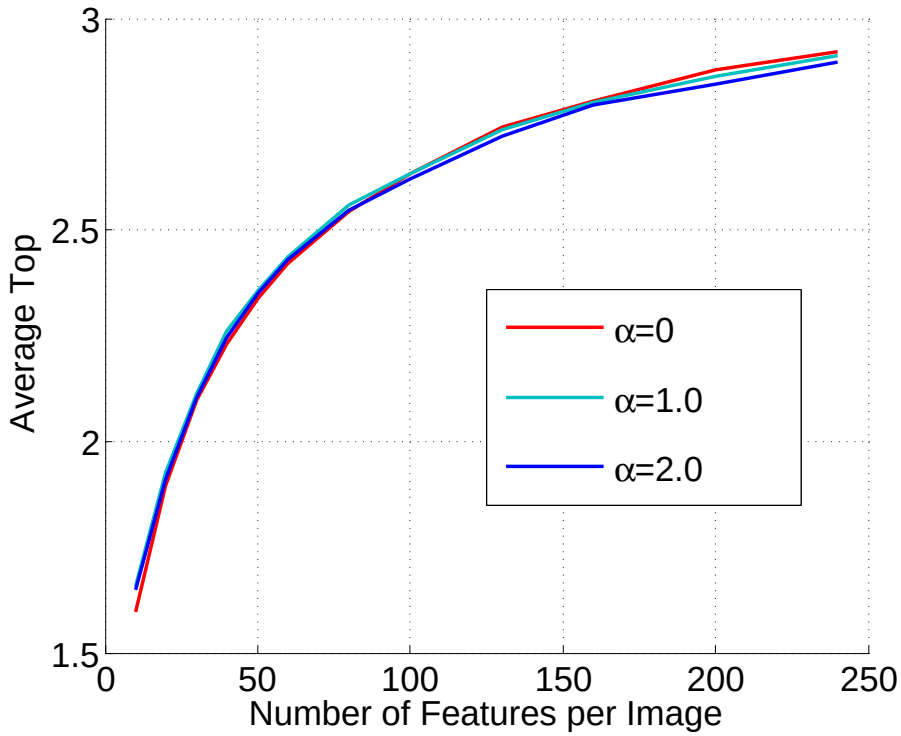


Figure 5.5: Effect of α on the Average Top Score.

In Table 5.3 we list the average number of hits and retrieval time per image using different number of features to represent each image. For a certain image, a hit happens when one of its visual words is found in another image. The number of hits directly affects the retrieval efficiency of the invert file indexing system. From the table we observe reduced number of features benefits the retrieval efficiency substantially. It is also interesting to note that at certain point where the number

Table 5.2: Number of features needed to achieve pre-defined Average Top on the UKB dataset

Target Average Top	1.8	2.2	2.4	2.6	2.9
Selection by Proposed Quality Metric	16.9	37.7	57.5	92.7	219.3
Selection by High Contrast	26.8	51.2	72.6	105.9	236.1
Selection by Large Scale	33.2	47.9	70.6	156.2	NA
Selection by the Quality Metric in [57]	29.2	74.6	118	186.6	NA

NA: Not Achievable in the experiments

The target Average Top is listed in the first row. The numbers of features needed to achieve the target average top are listed in the following rows.

of features is large enough, using more features actually degrades the matching accuracy.

It should be noted that the average top score we have obtained here is slightly worse than the numbers reported in other literatures. One possible reason is that we employ SIFT feature points (DoG point detector) that are only robust to scaling and

Table 5.3: Average number of hits and retrieval time

Number of features per image	50	100	200	300	500	700	1000
Number of hits per image	157.6	439.0	1265.4	2382.5	5425.0	9299.2	15159.4
Retrieval time per image(ms)	3.7	9.3	26.3	49.9	119.6	220.8	378.4
Average top score	2.33	2.63	2.87	2.96	3.02	3.01	2.96

rotation, whereas more sophisticated affine invariant interest point detectors are used in other literatures. Techniques that can improve the retrieval performance, such as Approximate K-means [48], soft visual word assignment [49] and query expansion [18], are not adopted in our current implementation. However, our experiments suffice to show the effectiveness of the proposed feature quality metric as a way of identifying and selecting useful features.

The Oxford Buildings Dataset

The Oxford Buildings Dataset [48] consists of 5062 images collected from Flickr by searching for Oxford landmarks. The dataset has been manually annotated to generate a comprehensive ground truth for 11 different landmarks, each represented by 5 queries. To evaluate retrieval performance, we use the Average Precision (AP), defined as the area under the precision-recall curve. Precision is the ratio of the number of retrieved true positive images to the total number of retrieved images. Recall is defined as the ratio of the number of retrieved true positive images to the the total number of true positive images in the gallery. The perfect precision-recall curve has precision 1 over all recall values, which corresponds to AP of 1. The AP score is computed for each of the 55 query images in the dataset. The average of these AP scores, termed as Mean Average Precision (mAP), is used to evaluate the overall performance on this dataset.

We use similar settings to the UKB dataset. A visual word vocabulary is obtained using the HKM method with 1 million cluster centers. Each query and

training image is represented by N selected SIFT features, and these features are mapped to visual words according to the vocabulary. In Figure 5.6, we plot the mAP scores for different values of N. Feature selection by the proposed quality metric again performs the best among all the methods that we have tested. To achieve the same mAP score, fewer features are needed by the proposed scheme than the selection by large scale or high contrast method, as can be seen from Table 5.4.

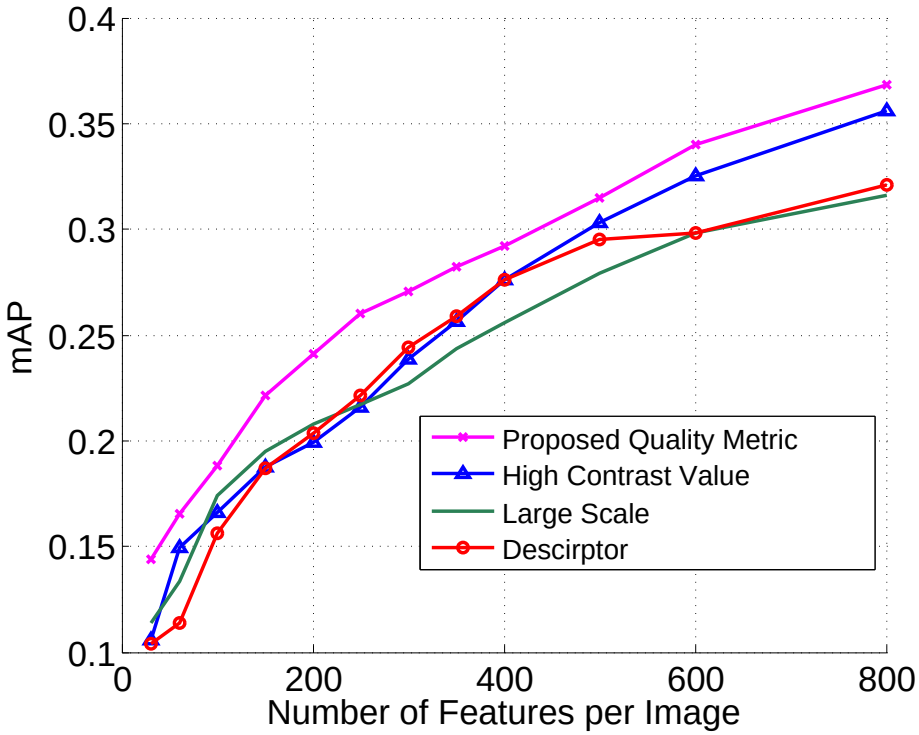


Figure 5.6: The mAP scores on the Oxford Buildings dataset using different feature selection schemes.

Table 5.4: Number of Features Needed to achieve Pre-defined mAP on the Oxford Dataset

Target mAP	0.15	0.2	0.25	0.3	0.35
Selection by Proposed Quality Metric	39.9	117.1	222.9	433.2	674.4
Selection by High Contrast	60	200.7	332.2	487.8	760.6
Selection by Large Scale	76	166.6	372.3	616.3	NA
Selection by the Quality Metric in [57]	94.8	86.4	320	612.5	NA

NA: Not Achievable in the Experiments

The INRIA Holidays Dataset

The INRIA Holidays dataset contains a set of personal holiday photos of mainly outdoor scenes captured by the authors of [34]. The dataset contains 1,491 images divided into 500 groups, each of which represents a distinct scene or object. In the experiment, we use the first image of every image group as query images and evaluate the mean average precision as defined in Sec. 5.3.2. As this dataset is relatively small, the retrieval is performed with nearest neighbor match instead of using the BoW approach. N SIFT features are selected from every image to form a feature gallery. When an image is used as a query, we take each of its N features and find the 3 nearest neighbors among the feature gallery. The images that contain at least one of the 3 nearest features receive one vote. After considering all the query features, the images with the most votes are regarded as top matches. Approximate Nearest Neighbor (ANN) techniques [8, 33, 41] are adopted to accelerate the nearest

neighbor search. Using the implementation provided by [41], we can speed the nearest neighbor search by as many as 50 times while guaranteeing that the searching result is correct with a probability above 95%. As illustrated in Figure 5.7 and Table 5.5, the proposed feature selection scheme continue to outperform the other methods.

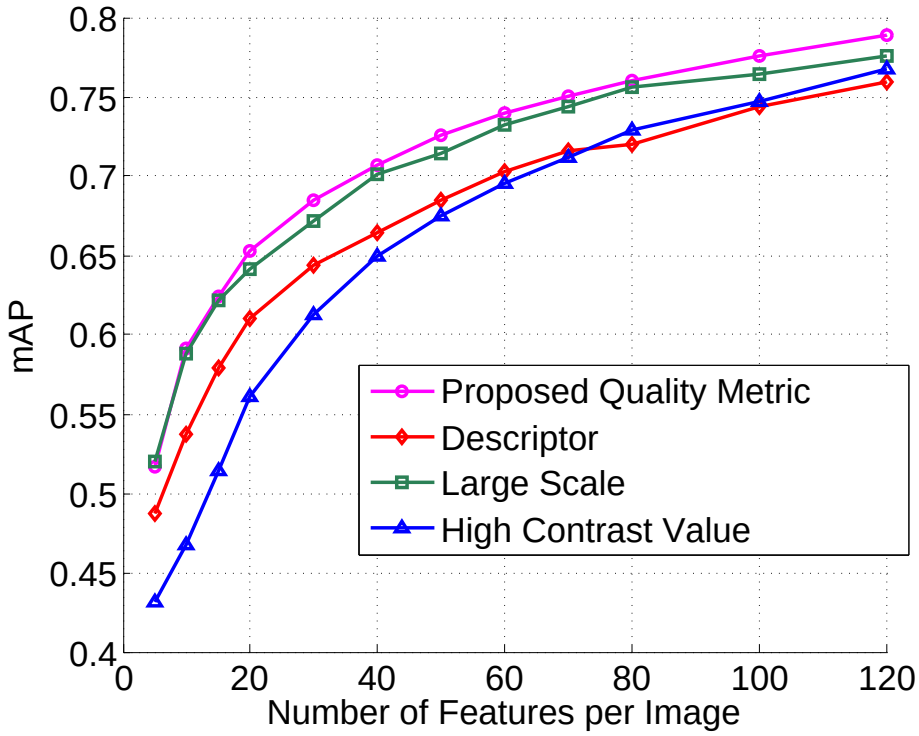


Figure 5.7: The mAP scores on the INRIA Holiday dataset using different feature selection schemes.

For mobile visual search applications, the size of the feature data transmission over the network between client and server can be further reduced by applying feature compression techniques [15,16,26] to the selected features. Here, we conduct experiments to reduce the feature dimensionality via Principle Component Analysis

Table 5.5: Number of Features Needed to achieve Pre-defined mAP on the INRIA Holiday Dataset

Target mAP	0.55	0.60	0.65	0.70	0.75
Selection by proposed quality metric	7.2	11.2	19.4	36.8	69.7
Selection by high contrast	18.8	28.7	40.3	63.7	103.7
Selection by large scale	7.2	11.8	22.8	40.6	75
Selection by the quality metric in [57]	11.7	18.9	33.4	59	108.2

(PCA). The eigenvectors are learned from all features in the dataset, and the features selected by the proposed quality metric is projected to M eigenvectors with the largest eigenvalues, resulting in M-dimension PCA features. Each dimension of the features is then quantized to 8 bits. These PCA features are used for nearest neighbor match. Figure 5.8 shows the mAP scores achieved with various feature size, from which we observe the efficiency is improved with PCA on the original SIFT features.

The proposed feature selection scheme using the quality metric performs consistently better than other empirical schemes on all three benchmark datasets that we have tested on. Each of these three datasets covers images of distinct and complementary characteristics. For example, the UKB contains images of mainly small objects, texture contents, and CD covers, and we note that features with high contrast are more informative than those of large scale. For the INRIA Holiday dataset that contains images of mostly outdoor scenes, features of large scale are more use-

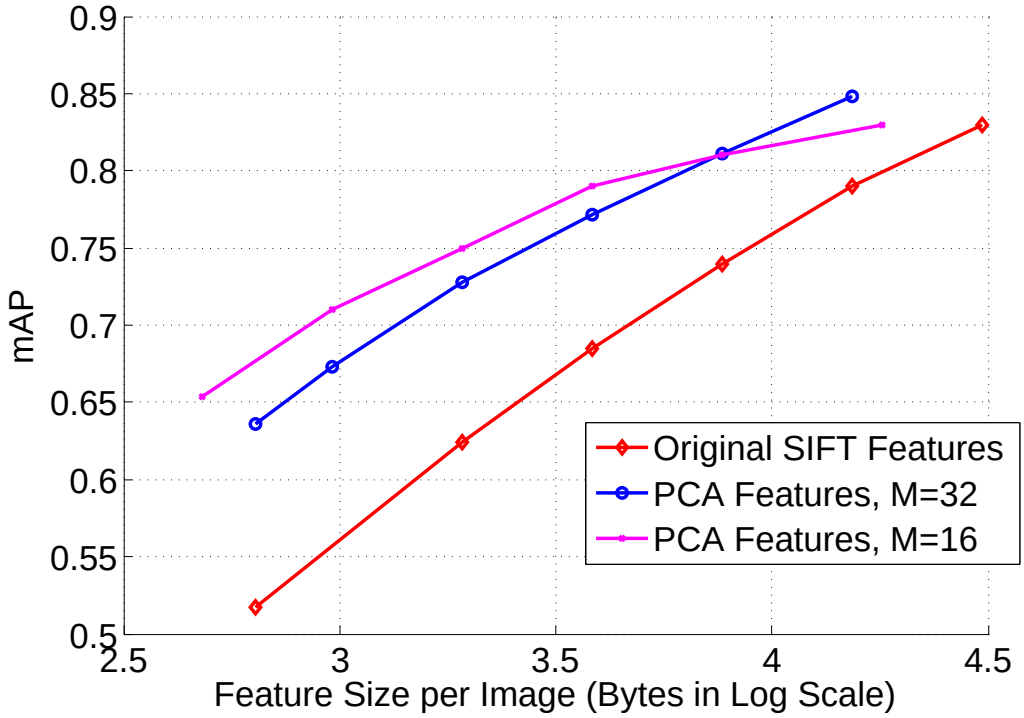


Figure 5.8: The mAP scores on the INRIA Holiday dataset as a function of feature size.

ful than high-contrast features. Our feature quality metric considers all the scale, contrast, and descriptor information, so it offers a more stable and effective method of selecting useful features for a broad variety of image contents.

5.4 Discussions

5.4.1 Examining the Quality Metric Kernel

To gain a better understanding of the characteristics of SIFT features, we look closely at the quality metric kernel. The kernel is essentially a 3-dimensional matrix, with each entry representing the robustness and discriminability of the point

in the feature space indexed by its scale, contrast, and descriptor. It is enlightening to examine the marginal quality metric, i.e., the quality score as a function of only one of the feature properties (scale, contrast, or descriptor). With the 3-dimensional kernel as a joint quality metric, we can compute the marginal of one feature property by averaging across the other two. In Figure 5.9 we demonstrate the quality score as functions of scale and contrast, respectively. The quality score reaches its peak at the scale near 10-20. Then, it starts to decrease significantly as the scale increases. However, features of extremely large scale (larger than 20) are rare (as can be seen from the CDF curve). So, although selecting features of the largest scale is not optimal, it still works much better than random selection. Similar rules apply for the key point contrast. As the contrast increases, the quality score first improves to its peak at around 20-25, then decreases.

Figure 5.10 provides some examples of visual words of high quality score (top row) and low quality score (bottom row). We observe that the descriptors with high scores generally have multiple significant gradient orientations, while those with low scores have a single outstanding peak along the dominant direction. This is correspond with the intuition that the former corresponds to image patches at complex structures with more discrimination power, and the latter corresponds to image patches around simple line structures, which are not as useful to distinguish different objects.

5.4.2 Generalization To Other Local Image Features

Although we demonstrated the feature quality evaluation scheme with SIFT features, we expect the methodology to apply to other image local features, such as Speeded Up Robust Features (SURF) [11] and Gradient Location and Orientation Histogram (GLOH) [40], as well. We demonstrate briefly the results we obtained with SURF. The key point detector in the SURF algorithm is an integer approximation to the determinant of Hessian blob detector. The descriptor is a 64 dimension vector comprising the statistics calculated from the wavelet responses of the 4 4-by-4 regions near the key point.

We repeat the feature quality kernel training process as we employed for the SIFT features, and conduct experiments on the INRIA Holiday dataset. We represent each image in the dataset with N SURF features selected with different strategies. The retrieval scores (mAP) as a function of N are illustrated in Fig. 5.11. The selection scheme based on the proposed feature quality metric outperforms the other empirical schemes, similar to SIFT features.

5.5 Chapter Summary

In this chapter, we have focused on local image feature selection for efficient visual matching. Local image feature presents a powerful tool to find point correspondences between multiple images of the same object or scene. Many of the most successful solutions for image registration and content-based image retrieval are built on the use of local image features. The increase of the image resolution

and the growth of the scale of image database may lead to the explosion of the number of image features. Feature selection becomes important in order to improve the feature matching efficiency.

To solve the problem of feature selection, we have presented a quality evaluation method for SIFT features. Our approach is built upon a quality metric kernel, which is essentially a 3-dimensional matrix indexed by SIFT feature's quantized scale, contrast and descriptor. In the training phase, a set of training images covering various scene types are collected. We conduct feature matching between the training images and their synthetically transformed versions. The matching results are aggregated to the metric kernel, so every entry of the kernel reflects how the features belonging to the corresponding scale-contrast-descriptor feature space bin behave statistically. The ratio of the number of correct matches over the total number of features in each bin is calculated as the quality score for the corresponding bin. In the application phase, a given feature's quality score is computed based on the its affinity to the feature space bins in the kernel and the quality scores of the bins. The proposed approach is tested on 3 benchmark datasets for large scale content-based image retrieval. Feature selection according to the proposed quality score is shown to perform better than the empirical methods, such as selecting features with largest scales and highest contrast values.

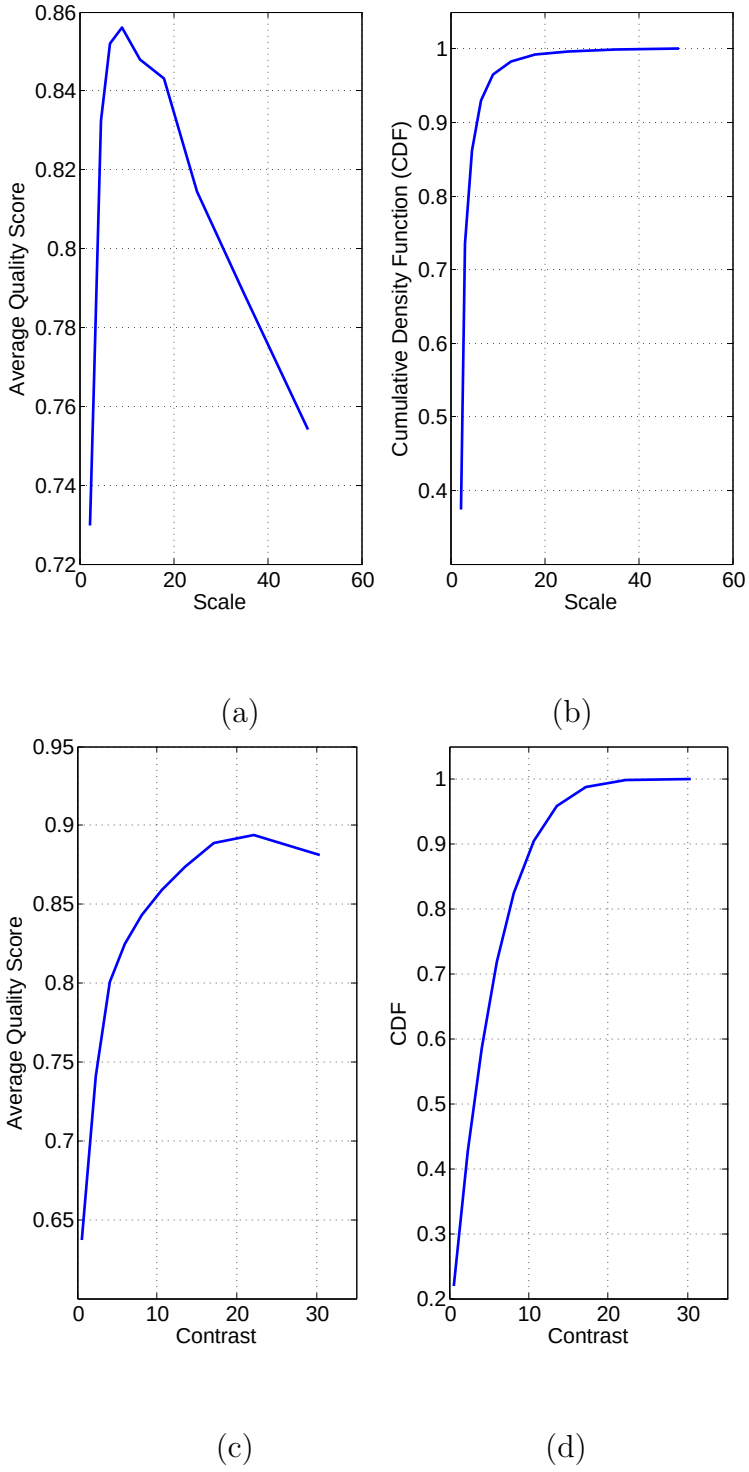


Figure 5.9: Analyzing the quality score as a function of scale and contrast value, respectively. (a) shows the quality score of features at different scales; (b) shows the cumulative distribution function (CDF) of feature scale; (c) shows the quality score of features with different contrast values; (d) shows the CDF of feature contrast value

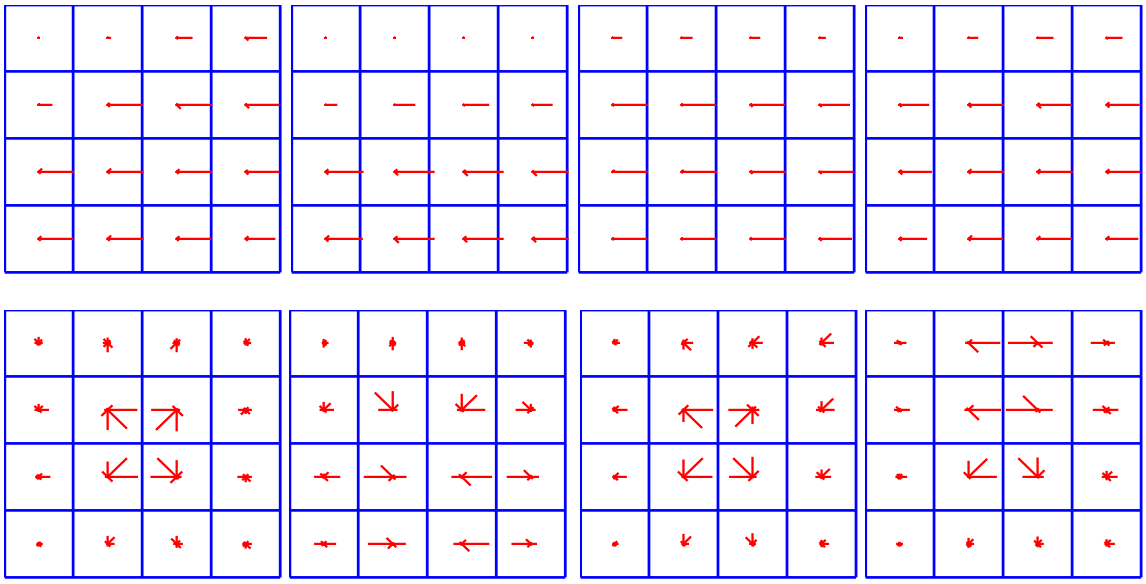


Figure 5.10: Visual words with low quality scores (top) and high quality scores (bottom).

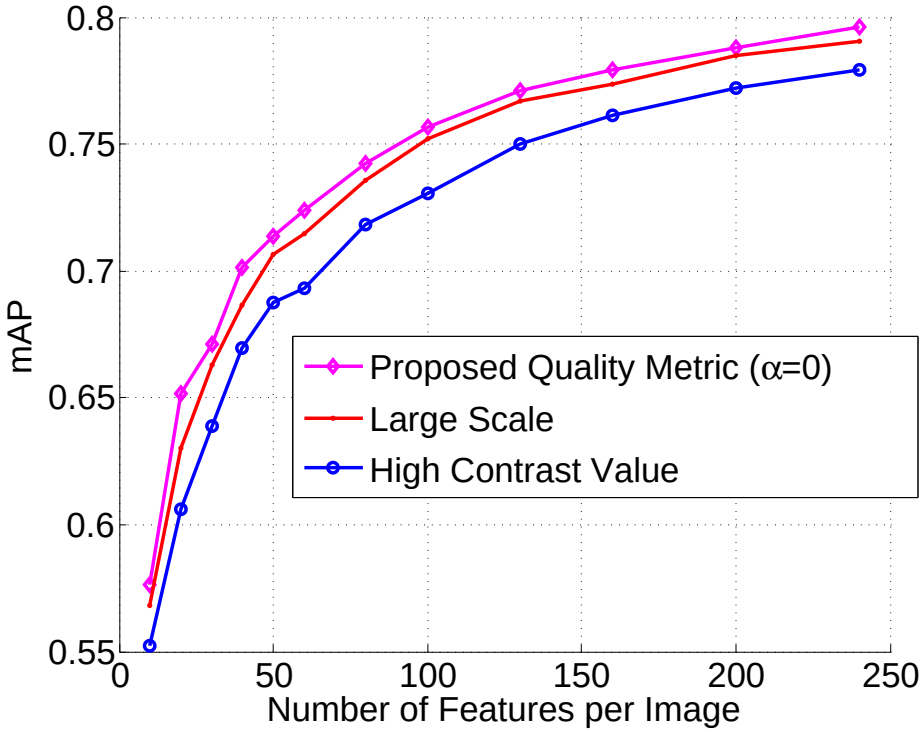


Figure 5.11: The mAP scores on the INRIA Holiday dataset using different SURF feature selection schemes.

Chapter 6

Conclusions and Future Perspectives

In this dissertation, we have considered the temporal and spatial alignment of multimedia signals. For temporal alignment, we propose a new modality for audio and video synchronization by exploiting the electric network frequency (ENF) signal naturally embedded in multimedia recordings. The value of the ENF signal fluctuates randomly around its nominal value over time, and the variation patterns remain consistent within the same power grid, even at distant locations. Synchronization of audio and video recordings can therefore be achieved by matching and aligning their embedded ENF signals. The ENF based method does not rely on the audio and visual contents in the multiple recordings to be synchronized. This property provides a strong potential to address such difficult scenarios that are intractable by existing methods. Taking video synchronization for example, the conventional approaches based on visual cues do not work well in situations with arbitrary camera motion or with insufficient view overlap, while the ENF based method is not affected by these

adverse conditions. Additionally, extracting and aligning ENF signals may be more effective computationally than the approaches that rely on computer vision and/or extensive learning, so more (or longer) recordings could be efficiently processed. It can also be easily generalized to synchronize multiple pieces of recordings. To the best of our knowledge, our work is the first in the endeavor to exploit the ENF for multimedia signal synchronization.

The extraction of ENF signals from multimedia recordings is extensively studied in the this dissertation, especially for visual recordings. ENF signals may be extracted from the soundtracks of the video recordings, as well as the image sequences if the video captures the subtle flickering of lightings. Extracting the weak ENF signal from image sequences presents a challenging task. The temporal sampling rate of visual recordings is generally too low for estimating the ENF signal that may appear at harmonics of 50 or 60 Hz. The ENF traces in video signals are relatively weak, and may be easily distorted by object and camera motions. Along with the direction of previous work, we have performed a further study exploiting the rolling shutter to extract ENF traces from video recordings. We model and analyze the rolling shutter mechanism with a filter bank using multirate signal processing theory. We then extend the scope of extracting ENF traces from videos of still scenes to those containing motions, which is a challenging problem and has never been formally attempted. We have proposed several techniques to overcome the difficulties of extracting the ENF signal from image sequences, such as the low sampling rate, object motions in the scene, camera motions, and brightness change.

We also address several challenges unique to the ENF analysis of recaptured

audio recordings, such as digitized historical recordings. Multiple ENF traces may exist in recaptured audio recordings. Traditional estimation methods may fail when the ENF signals corresponding to different time instances interfere with each other. A decorrelation based method is proposed to extract multiple ENF signals from recaptured audio recordings in a sequential manner. The ENF signal in a digitized analog recording may also suffer from the drifting effect, which refers to the phenomenon that the value of the ENF signal deviates severely from its original value. The mechanical rolling speed of analog tape recorder and player is usually not well stabilized. When an analog tape recording is digitized, the rolling speed of the tape is likely to differ from that of the creation of the tape. As a result, a speed offset is induced in the digitized version of the audio, causing the drifting effect. We have proposed to detect and correct such speed errors using the ENF signal as a reference and utilizing digital signal processing techniques. The proposed method is used to restore several digitized historical audio recordings of the NASA Apollo 11 Mission.

Local image features are widely used to detect common objects among multiple images and to obtain the spatial alignment of the images. In the local feature framework, interest points are first selected as distinctive and robust points in the image by a key point detector. Next, a robust feature descriptor is generated using the information within the neighborhood of the interest point. Since the images may have gone through different lighting and viewing conditions, desirable features should be robust to various distortions.

Many of the most successful solutions for image registration and content-based image retrieval are built on the use of local image features. The increase of the image

resolution and the growth of the scale of image database may lead to the explosion of the number of image features. Feature selection becomes important in order to improve the feature matching efficiency. To solve the problem of feature selection, we have presented a quality evaluation method for SIFT features. Our approach is built upon a quality metric kernel, which is essentially a 3-dimensional matrix indexed by SIFT feature's quantized scale, contrast, and descriptor. In the training phase, a set of training images covering various scene types are collected. We perform feature matching between the training images and their synthetically transformed versions. The matching results are aggregated to the metric kernel, so every entry of the kernel reflects how the features belonging to the corresponding scale-contrast-descriptor feature space bin behave statistically. The ratio of the number of correct matches over the total number of features in each bin is calculated as the quality score for the corresponding bin. In the application phase, a given feature's quality score is computed based on the its affinity to the feature space bins in the kernel and the quality scores of the bins. The proposed approach is tested on 3 benchmark datasets for large scale content-based image retrieval, and is shown to outperform the empirical methods.

Several interesting research topics along the directions of the work in this dissertation can be further explored. The ENF-based approach emerges as a new modality for audio and video synchronization. Unlike most conventional methods taht extract and match certain audio and visual features from the contents of the recordings, it utilizes the embedded ENF traces that are usually considered as noise. The ENF-based approach and the audio/visual cue-based approach may therefore

complement each other and may be combined to solve the problem of multimedia synchronization more effectively. For example, the ENF-based method can first be applied to obtain synchronization of two video clips efficiently with frame-level accuracy. The synchronization then can be refined by certain visual cue based-method to a sub-frame accuracy with more computation. Some video clips may consist of segments that are in different conditions for various synchronization methods. For a video clip, we may choose the best synchronization method for each segment, and combine these results of the segments to reach the final lag estimation for the whole clip. Video synchronization systems that exploit multiple modalities can be explored in the future.

The estimation of the ENF signal is crucial for the proposed multimedia synchronization method that relies on matching the embedded ENF signals. In the dissertation, we use the short time Fourier transform (STFT) based-method to extract ENF signals from audio and visual recordings due to its efficiency and simplicity. This method divides a source signal into possibly overlapping frames of small durations. Within every frame, the signal can be regarded as wide-sense stationary, and each of the frames undergoes Fourier analysis respectively. For ENF estimation, we apply STFT to a source signal that contains ENF traces, and find the peak frequency within a certain range near the nominal value or the harmonics in each frame. The values of the peak frequency from all the frames are concatenated to form the estimated ENF signal.

Several challenges need to be addressed in order to improve the synchronization accuracy. The STFT based-method works well when the signal to noise ratio (SNR)

of the ENF signal is fairly high. However, when the noise distortion is strong, the estimation performance degrades severely. Another problem is the temporal resolution of the frequency estimation. The length of the frame needs to be large enough to ensure reliable estimation, which limits the resolution of the estimate. The resolution of frequency estimation is critical to the multimedia signal synchronization accuracy. How to improve the frequency estimation accuracy in the case of low SNR and to increase the resolution is therefore an important research topic.

With the image feature selection scheme proposed in this dissertation, we can identify image features that are statistically more likely to be useful for describing visual structures. By retaining only these features, the transmission load and the storage cost for image features can be lowered. The matching accuracy can be potentially improved as well. However, the computation for generating the features is not reduced in any way, because the feature selection is performed after all feature candidates have been generated. An interesting direction to explore is to study whether feature selection can be carried out earlier and incorporated into the feature generation process. Much of the time for feature generation is spent on calculating the feature descriptor vector. Similar to the approach purposed in this dissertation, for a interest point, we can try to predict its robustness and discriminability using such information as its scale, contrast value, and some other features abstracted from its neighbor pixels. Then the feature descriptors are calculated only for those interest points that obtain a high quality score. The saving in descriptor calculation may improve feature generation efficiency substantially.

Bibliography

[1] Apollo 13 audio recordings, https://archive.org/details/Apollo13Audio.

[2] The first men on the moon: The apollo 11 lunar landing, http://www.firstmenonthemoon.com/.

[3] *Google Goggles*, http://www.google.com/mobile/goggles.

[4] *Nokia Point and Find*, http://www.pointandfind.nokia.com.

[5] *SnapTell*, http://www.snaptell.com.

[6] A. Agarwala, K. C. Zheng, C. Pal, M. Agrawala, M. Cohen, B. Curless, D. Salesin, and R. Szeliski. Panoramic video textures. In *SIGGRAPH*, 2005.

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

[8] S. Arya, D.M. Mount, N.S. Netanyahu, R. Silverman, and A.Y. Wu. An optimal algorithm for approximate nearest neighbor searching. *Journal of ACM*, 45:891–923, 1998.

[9] Miller Center at Univ. of Virginia. Presidential recordings, http://millercenter.org/scripps/archive/presidentialrecordings/kennedy.

[10] P. K. Baheti, A. Swaminathan, M. Chari, S. Diaz, and S. Grzechnik. Information-theoretic database building and querying for mobile augmented reality applications. In *IEEE International Symposium on Mixed and Augmented Reality*, 2011.

[11] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded up robust features. In *ECCV*, 2006.

[12] D. Bykhovsky and A. Cohen. Electrical network frequency (ENF) maximum-liklihood estimation via a multitone harmonic model. *IEEE Transactions on Information Forensics and Security*, 8(5), May 2013.

[13] Y. Caspi and M. Irani. A step towards sequence -to-sequence alignment. In *Int'l Conf. Comput. Vision & Pattern Recognition*, 2000.

[14] Y. Caspi, D. Simakov, and M. Irani. Feature-based sequence-to-sequence matching. *Int'l J. Comput. Vision*, 68(1), Jun. 2006.

[15] V. Chandrasekhar, M. Makar, G. Takacs, D. Chen, S. Tsai, N. Cheung, R. Grzeszczuk, Y. Reznik, and B. Girod. Survey of SIFT compression schemes. In *Proc. Visual Communication and Image Processing*, 2009.

[16] V. Chandrasekhar, G. Takacs, D. M. Chen, S. S. Tsai, J. P. Singh, and B. Girod. Transform coding of image feature descriptors. In *Proc. Visual Communication and Image Processing*, 2009.

[17] W. H. Chuang, R. Garg, and M. Wu. How secure are power network signature based time stamps? In *ACM Conf. on Computer and Communication Security*, Oct. 2012.

[18] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *CVPR*, 2007.

[19] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham Mysore, Fredo Durand, and William T. Freeman. The visual microphone: Passive recovery of sound from video. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 33(4):79:1–79:10, 2014.

[20] M. El-Saban, A. Kaheel, and M. Refaat. Stitching videos streamed by mobile phones in real-time. In *ACM Multimedia*, Oct. 2009.

[21] G. Fritz, C. Seifert, and L. Paletta. A mobile vision system for urban detection with informative local descriptors. In *Proc. of the Fourth IEEE Int. Conf. on Computer Vision Systems*, Jan. 2006.

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

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

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

[25] R. Garg, A. L. Varna, and M. Wu. Modeling and analysis of electric network frequency signal for timestamp verification. In *IEEE International Workshop on Information Forensics and Security*, Dec. 2012.

[26] B. Girod, V. Chandrasekhar, D. Chen, N.-M. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S. Tsai, and R. Vedantham. Mobiel visual search. *IEEE Signal Processing Magazine*, July 2011.

[27] C. Grigoras. Applications of ENF criterion in forensics: Audio, video, computer and telecommunication analysis. *Foresnsic Science International*, 167(2-3):136–145, April 2007.

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

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

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

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

[32] M. Huijbregtse and Z. Geradts. Using the ENF criterion for determining the time of recording of short digital audio recordings. In *International Workshop on Computational Forensics (IWCF)*, Aug. 2009.

[33] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the 30th Annual ACM Symposium on Theory of Computing*, 1998.

[34] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, 2008.

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

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

[37] Y. Liu, J. Chai, B. Greene, R. Conners, and Y. Liu. A study of the accuracy and precision of quadratic frequency interpolation for ENF estimation. In *Proceedings of the AES 46th international conference*, June 2012.

[38] D. Lowe. Distinctive image features from scale invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[39] Wenjun Lu, Avinash L. Varna, and Min Wu. Forensic hash for multimedia information. In *Proc. of SPIE Media Forensics and Security*, 2010.

[40] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27:1615–1630, October 2005.

[41] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *International conference on computer vision and applications*, 2009.

[42] N. Naikal, A. Y. Yang, and S. S. Sastry. Informative feature selection for object recognition via sparse PCA. In *ICCV*, Nov. 2011.

[43] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2006.

[44] O. Ojowu, J. Karlsson, J. Li, and Y. Liu. ENF extraction from digital recordings using adaptive techniques and frequency tracking. *IEEE Transactions on Information Forensics and Security*, 7(4):1330–1338, August 2012.

[45] A. V. Oppenheim and R. W. Schafer. *Discrete-Time Signal Processing*. Prentice Hall, 2009.

[46] F. Padua, R. Carceroni, G. Santos, and K. Kutulakos. Linear sequence-to-sequence alignment. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(2), Feb. 2010.

[47] M. Pawig, G. Enzner, and P. Vary. Adaptive sampling rate correction for acoustic echo control in Voice-Over-IP. *IEEE Transactions on Signal Processing*, 58(1), May 2013.

[48] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[49] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, 2008.

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

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

[52] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *IJCV*, 37(2):151 – 172, 2000.

[53] E. Shechtman, Y. Caspi, and M. Irani. Space-time super-resolution. *IEEE Trans. Pattern Anal. Mach. Intell.*, Apr. 2005.

[54] P. Shrestha, H. Weda, M. Barbieri, and D. Sekulovski. Synchronization of multiple video recordings based on still camera flashes. In *ACM Multimedia*, 2006.

[55] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. *International Conference on Computer Vision*, 2003.

[56] G. P. Stein. Tracking from multiple view points: self-calibration of space and time. In *Int'l Conf. Comput. Vision & Pattern Recognition*, 1999.

[57] H. Su, W-H. Chuang, W. Lu, and M. Wu. Evaluating the quality of individual SIFT features. In *ICIP*, 2012.

[58] H. Su, R. Garg, A. Hajj-Ahmad, and M. Wu. ENF analysis on recaptured audio recordings. In *IEEE Int. Conf. Acoustics, Speech, & Signal Process. (ICASSP)*, May 2013.

[59] H. Su, A. Hajj-Ahmad, R. Garg, and M. Wu. Exploring rolling shutter for ENF extraction from video. In *ICIP*, 2014.

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

[61] H. Su, C.W. Wong, A. Hajj-Ahmad, R. Garg, and M. Wu. ENF signal induced by power grid: a new modality for video synchronization. In *ACM MM Workshop on Immersive Media Experiences*, 2014.

[62] P. Turcot and D. Lowe. Better matching with fewer features: The selection of useful features in large database recognition problems. In *ICCV Workshop on Energent Issues in Large Amounts of Visual Data*, 2009.

[63] T. Tuytelaars and L. Van Gool. Synchronizing video sequences. In *Int'l Conf. Comput. Vision & Pattern Recognition*, 2004.

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

[65] A. Vedaldi and B. Fulkerson. VLFeat - an open and portable library of computer vision algorithms. *ACM Multimedia*, 2010.