공학박사 학위논문

Efficient Control System for Smartphone Sensor-based Applications

스마트폰 센서 기반 어플리케이션의 효율적인 동작을 위한 시스템 설계

2023년 2월

서울대학교 대학원

전기정보공학부

곽 철 영

공학박사 학위논문

Efficient Control System for Smartphone Sensor-based Applications

스마트폰 센서 기반 어플리케이션의 효율적인 동작을 위한 시스템 설계

2023년 2월

서울대학교 대학원

전기정보공학부

곽 철 영

Efficient Control System for Smartphone Sensor-based Applications

지도교수 박 세 웅

이 논문을 공학박사 학위논문으로 제출함

2023년 2월

서울대학교 대학원

전기정보공학부

곽 철 영

곽철영의 공학박사 학위 논문을 인준함

2022년 10월

위 원 장: 최 완 (인)

부위원장: 박 세 웅 (인)

위 원: 이 경 한 (인)

위 원: 김 형 신 (인)

위 원: 백 정 엽 (인)

Abstract

With the advent of smartphones, mobile devices are equipped with various sensors. Computational capabilities and a variety of sensors enable many new things on mobile devices. In recent years, many researchers have tried to expand the boundary of smartphone applications with previously unavailable media types. For example, acoustic communication using the microphone and speaker of a smart device has been one of the hottest wireless technologies in recent years. Mobile text spotting, interaction with text through a camera sensor, is also one of the active research topics in academia and industry. For the operation of these new attempts in mobile devices, efficiency and practicality are essential issues.

In this dissertation, we propose two systems that enable an efficient operation of various smartphone sensor-based applications: (i) No Entry: Anti-Noise Energy Detector for Chirp-Based Acoustic Communication and (ii) Cameleon: Intelligent Camera Sensor System for Text-spotting Oriented Operation in Mobile Devices.

First, No Entry is a novel energy detector (ED) for chirp-based acoustic communication systems. No Entry avoids not only high-energy noises but also a different modulation-based acoustic signal by utilizing the frequency sweeping characteristic of chirp signals. We implement prototype Android applications to evaluate detection accuracy and power consumption. Compared with the state-of-the-art schemes, No Entry reduces energy consumption by 30% while achieving a greater detection performance.

Second, we propose a camera sensor control system for deep learning applications in mobile devices. While deep learning model benchmark dataset results promise outstanding performance, in reality, the quality of sensor data has a significant impact on the performance. We design an intelligent text-spotting oriented camera sensor control system. Unlike the default camera operation that produces images good for the human eye, the proposed camera sensor control system controls the camera sensor

optimized for the text spotting task. We implement and validate our design through extensive experiments. Compared to the traditional camera pipeline, Cameleon dramatically recovers performance degradation and maximizes the text-spotting model's performance.

In summary, we propose systems that enable the efficient operation of smartphone sensor-based applications. We implement two systems on commercial smartphones. We also verify the performance through extensive real-world experiments. Through this research, we take a step to expand the boundary of smartphones' potential.

keywords: smartphone, smatphone sensor, mobile application, mobile deeplearning
student number: 2015-20885

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Recently, with the rapid growth of mobile devices, there have been many attempts to utilize sensors in mobile devices. Various sensors and more powerful computational capabilities of mobile devices have enabled different types of interactions with various types of media in smartphones. In recent years, many researchers have tried to expand the boundary of smartphone interaction with previously unavailable media types.

We introduce two types of smartphone applications operating with different mobile device sensors based on microphone and speaker sensor and camera sensor, respectively. We briefly introduce each system and point out existing problems depreciating their practicality. In this dissertation, we present how to solve the existing limitations of each system and suggest performance improvements.

Acoustic communication: Acoustic communication using microphones and speakers of smart devices is one of the most spotlighted wireless technologies in recent years. Acoustic communication in mobile device utilizes near-ultrasound frequency band that belongs to audible frequency band but people hardly hear. In particular, chirp-based acoustic communication is widely adopted for smart device applications because of its robustness to frequency selectivity. Acoustic communication is widely

used in localization, communication, etc. Since these types of applications run in the background, the power consumption is an essential issue. In general, wireless communication save energy using energy detectors (EDs), which determine the existence of a signal based on energy level. Acoustic coomunication can also reduce power consumption by working only when a valid signal exists. However, there exists lots of noises in near-ultrasound frequency band, and thus it is hard for conventional ED to distinguish them from noise. In order to solve this problem, we design a novel ED for chirp-based acoustic communication system to distinguish various noise and interference existing in near-ultrasound frequency.

Mobile text-spotting: Mobile text-spotting using camera sensor of mobile device is one of the active research subjects in both academics and industries close to real-life usage. While text-spotting has been developed dramatically with the rising of deep learning, none of the previous work consider the process of making input in mobile devices. People have focused on improving the performance of text-spotting model on a benchmark dataset. We attack the weak spot of deep learning models, it severly underperforms when the input is different to the trained dataset. Since mobile device camera is not optimized for text-spotting, an existing mobile device camera fails to producing good images for text-spotting in some environments. In order to solve this problem, we design a novel text-spotting oriented intelligent camera control system.

1.2 Main Contributions

1.2.1 No Entry: Anti-Noise Energy Detector for Chirp-Based Acoustic Communication

We propose a novel ED for chirp-based acoustic communication systems that overcomes existing limitations. We scrutinize the chirp signal, and find a distinct characteristics of chirps. We design an ED that find chirp-signal by checking the special key features of chirp. Since the merit of ED is computational simplicity, we design the

detection algorithm not to deteriorate the simplicity. The main contributions of this chapter are as follows.

- We propose No Entry, a novel ED that can avoid not only high-energy noise but also high-energy interference by utilizing the frequency sweeping characteristic of chirp signals.
- Detection accuracy of No Entry shows that true positive (TP) rate is more than 90% when false positive (FP) rate is 1% even with severe interference.
- The power consumption of No Entry is measured using a prototype Android application and Monsoon power monitor. No Entry reduces energy consumption by about 30% compared with the state-of-the-art scheme.

1.2.2 Cameleon: Intelligent Camera Sensor System for Text-spotting Oriented Operation in Mobile Devices

We aim to propose a system to control cameras for text-spotting-oriented operations, which reduce the performance degradation from conventional camera operation. The main contributions of this chapter are as follows.

- We verify the operation of the mobile device camera is not optimized to work well with a text-spotting network. We show that the text-spotting network suffers from a performance drop in accordance with changes in the environment of the input images.
- We propose Cameleon, a novel intelligent camera sensor control system that controls the camera exposure to fit the image for the text-spotting network in various environments.
- We validate our design through extensive experiments over the various places and objects. We verify that Cameleon generates better images for text-spotting application than that of conventional camera systems.

1.3 Organization of the Dissertation

The rest of the dissertation is organized as follows.

Chapter 2 presents No Entry, a novel ED for chirp-based acoustic communication. No Entry overcomes the existing limitation of conventional ED coming from the ambient noise in everyday life. We present the overview and the detailed process of No Entry and suggest extensive evaluation in real-world experiments.

In Chapter 3, we present a text-spotting-oriented intelligent camera control system called Cameleon. The main philosophy in desining Cameleon is to control mobile device camera in a way to fit text-spotting. We explain Cameleon in detail and suggest evaluations through comprehensive real world experiments.

Finally, Chapter 4 concludes the dissertation with the summary of contributions and discussion of the future work.

Chapter 2

No Entry: Anti-Noise Energy Detector for Chirp-Based Acoustic Communication

2.1 Introduction

Smart mobile devices are no longer special things to modern people. As smart devices become common, they play several roles. For example, they play the role of credit cards, coupon books, and even gaming consoles. Accordingly, various types of wireless communication system have been proposed to deal with several applications. An acoustic communication system using smart mobile devices is one of the most spotlighted wireless technologies.

Most acoustic communication systems using smart devices operate in the near-ultrasound frequency range, *i.e.,* 18–20 kHz. The lower frequency range is avoided to prevent any unwanted audibility by humans, and the higher frequency range is limited by the sampling rate of 44.1 kHz, which is mostly common in off-the-shelf smart devices. A lot of research utilizing the advantages of the sound has been proposed. Relatively slower speed of sound has been flourishing the indoor localization and motion tracking research [4,5]. In addition, the feature that near-ultrasound acoustic signal can be easily embedded in the music or video contents without user's perception attracts

the attention of the short-range data transmission services [6–9].

Acoustic communication exploits various digital modulation schemes, such as frequency-shift keying (FSK), orthogonal frequency-division multiplexing, chirp modulation, etc. In this chapter, we focus on chirp signal-based acoustic communication systems. Chirp signals have a clear advantage in that they are robust to frequency selectivity [7]. Thus, chirp signals are suitable for some background applications such as second screen service [8] and motion tracking [4], which requires robust signal transmission as the top priority.

Unlike foreground applications, which a user works on in person, background applications work continuously behind the scene often without being perceived by the user. Thus, background applications pose a risk of excessive power consumption without a user perception. To be specific, in most existing acoustic-based background applications, a receiver does not have any preliminary knowledge about a transmitter, and hence, the receiver has to periodically wake up and try to receive a signal. However, applications consume considerable energy in the receiving process. Applications can reduce energy consumption by detecting a signal preferentially and working only when a valid signal exists.

Energy detector (ED) is a type of such signal detection system. A conventional energy detection method measures *in-band energy*, *i.e.,* the energy conveyed in the frequency range used by the communication system, during a detection time [10]. However, if we use the conventional ED in acoustic communication, there are two problems. The first problem is caused by some types of noise, called ambient noise, which influences acoustic communication. Fig. 2.1 shows the power spectral density (PSD) plot of two different types of noise, *i.e.,* a conversation sound and a cough sound. It is shown that the power level of the conversation sound decreases sharply from the frequency of about 10 kHz. On the contrary, the cough sound keeps relatively high power level up to the frequency of 20 kHz. Due to the high power level, an ED can mistake the cough sound as a valid signal. Likewise, ambient noise refers to such

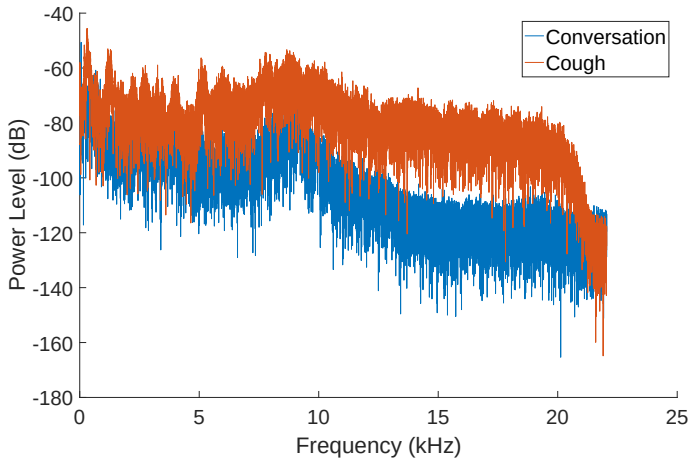


Figure 2.1: PSD plots of a conversation sound and a cough sound.

noise whose frequency components have a high power over the entire audio frequency, so that an ED can mistake ambient noise as a signal. Second, if there exists an interference, *i.e.,* a signal from different acoustic communication systems sharing in-band frequency range, ED can also mistake the interference as the signal of its own system.

In this chapter, we propose No Entry, a novel ED for chirp-based acoustic communication systems that goes beyond existing limitations. The main idea behind No Entry is not merely to sense the in-band energy level, but to verify that the signal has the frequency sweeping characteristic of chirp signals. Since the merit of ED is computational simplicity, we design the detection algorithm not to deteriorate the simplicity. The main contributions of this chapter are as follows.

1. We propose No Entry, a novel ED that can avoid not only high-energy noise but also high-energy interference by utilizing the frequency sweeping characteristic of chirp signals.

2. Detection accuracy of No Entry shows that true positive (TP) rate is more than 90% when false positive (FP) rate is 1% even with severe interference.

3. The power consumption of No Entry is measured using a prototype Android

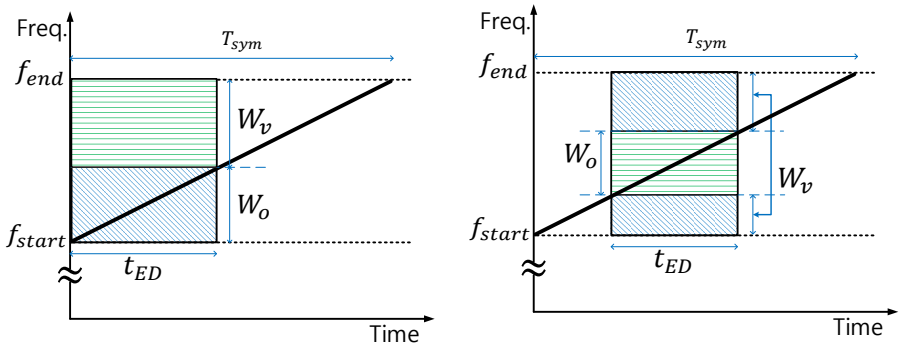
application and Monsoon power monitor. No Entry reduces energy consumption by about 30% compared with the state-of-the-art scheme.

The rest of this chapter is organized as follows. In Section 2.2 and Section 2.3, we describe related work and background, respectively. Section 2.4 presents the overall system of the proposed ED. We discuss several parameters related to implementation and analyze the computation complexity of the proposed method in Section 2.5. In Section 2.6, we evaluate the performance and conclude the chapter in Section 2.7.

2.2 Related Work

There have been many studies to deal with the signal detection in acoustic communication. The authors of [7,11] use *peak PSD ratio* of in-band, *e.g.,* 19.5–22 kHz frequency range, to out-of-band, *e.g.,* 16–18 kHz frequency range. The authors assume both in-band and out-of-band have similar noise levels when the signal from their acoustic communication system is absent. In this case, peak PSD ratio becomes low because the peak PSD of the in-band and that of the out-of-band are comparable. If the signal exists in the in-band, on the other hand, the peak PSD of the in-band is much greater than that of the out-of-band, so that the peak PSD ratio becomes high. However, this method has limitations. If there exists ambient noise that has higher energy in the in-band than in the out-of-band, peak PSD ratio becomes high thus causing a false positive error. Peak PSD ratio also has a trouble distinguishing different acoustic signals from different communication systems which use the frequency range nearby the in-band.

The authors of [8] consider the influence of ambient noise and propose *J-CS* algorithm utilizing the shape of a chirp signal correlation. J-CS is able to distinguish the non-chirp signals because it uses chirp signal's correlation results. However, even though J-CS is able to differentiate interference signals from the chirp signals, it has a limitation in terms of power consumption. To be specific, J-CS hardly reduces power



(a) Continuous case

(b) Circularly continuous case

Figure 2.2: Example of occupied/vacant frequency.

consumption because it determines the existence of the signal at the end of its receiving process.

The authors of [12] propose an energy-efficient acoustic communication system. They utilize an acoustic beacon signal to determine the presence of an acoustic signal before operating the entire process. A receiver wakes up periodically and verifies whether a beacon signal exists or not during short detection time. If the receiver detects the beacon signal, it tries to receive the signal. If the receiver does not detect the beacon signal, on the other hand, it goes to sleep. However, the authors do not specify the behavior of the beacon detection process in the chapter, as their main contribution is an accurate spatially-aware interaction.

2.3 Background

2.3.1 Chirp Signal

A chirp signal is a signal whose frequency sweeps over time, *i.e.,* the frequency increases or decreases with time. We exploit this frequency sweeping characteristic of chirp signals. Assume that we observe a chirp signal during a time duration shorter

than symbol duration (T_{sym}), *i.e.,* the *detection time*, notated by t_{ED}, is smaller than T_{sym}. Then, the signal sweeps not the entire in-band frequency range but a part of the in-band frequency range. We can divide the in-band frequency range into two groups. We define *occupied frequency range* (W_o) as the frequency range the signal sweeps during t_{ED}, and *vacant frequency range* (W_v) as the rest in-band frequency range excluding W_o. Fig. 2.2 shows an example of a chirp signal (a linearly-increasing straight line) of duration T_{sym} sweeping from f_{start} to f_{end}, along with t_{ED}, W_o, and W_v as well as their relationships. Even though the ratio between W_o and W_v would change depending on t_{ED} and chirp's sweeping rate, each of W_o and W_v is continuous or circularly continuous within the in-band frequency range. Circularly continuous means the end of the in-band frequency (f_{end}) is followed by the beginning of the in-band frequency (f_{start}), as shown in Fig. 2.2b. We utilize two attributes of a chirp to distinguish it from ambient noise, *i.e.,* 1) in-band frequency range consists of W_o and W_v, and 2) each of W_o and W_v is (circularly) continuous. We will present the details in the next section.

2.3.2 Noise Analysis

One of the biggest weaknesses of acoustic communication is its vulnerability to ambient noise. Compared to radio frequency wireless technologies, acoustic communication uses relatively low frequency range around kilohertz where noise can be easily generated by human activities. The authors of [8] describe the influence of ambient noise on acoustic communication. However, they do not survey how often ambient noise is generated. If ambient noise rarely appears in everyday life, handling ambient noise in the ED might rather be an unnecessary overhead.

We investigate how often ambient noise is generated in two different environments. We collect data in two offices representing quiet places and two cafes representing loud places. We record one hour per measurement and check how often ambient noise exceeds a certain energy threshold. To set the energy threshold, we select the appli-

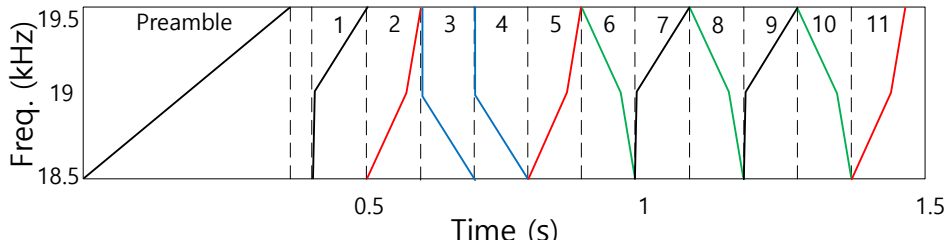


Figure 2.3: Chirp frame structure.

cation in [8], which is a chirp-based background second screen service application, as our target application. In the target application, the signal's frequency range, *i.e., in-band frequency range*, is 18.5–19.5 kHz. As shown in the Fig. 2.3, a frame consists of 368 ms-long up-chirp preamble, 40 ms-long post-preamble guard interval (GI), and 11 symbols with *symbol duration* (T_{sym}) of 96 ms each, making the overall packet duration 1463 ms [8]. We first generate and collect the chirp signals 10,000 samples in a quiet conference room environment using TV as transmitter. To specify the volume of the signal, we measure the sound pressure level and set the sound pressure level of the signal to 32 dBSPL, which is very low volume sound similar to the sound level in a quiet bedroom at night [13]. We then set the threshold as the average energy of the collected data. We measure the ratio of the number of samples exceeding the threshold to the whole samples for a detection time of 50, 100, and 200 ms. The measurement result is shown in Table 2.1.

The result in the offices, which are relatively quiet, is less than 5% because ambient noise is rarely generated in a quiet environment. Due to a fricative sound containing high in-band energy such as a door closing sound, however, ambient noise is sometimes generated as shown in the result in Office 1. On the other hand, the result in the cafes, which are relatively loud, shows that lots of samples pass the energy threshold. In cafes, noise is generated by the coffee machine sound as well as dragging chair or desk sound. These types of noise are common and frequently generated during a busy

Table 2.1: Measurement Results

Detection Time	Place			
	Cafe1	Cafe2	Office1	Office2
50 ms	46.1%	59.8%	2.8%	0.9%
100 ms	34.0%	45.3%	3.4%	0.7%
200 ms	28.9%	35.0%	4.0%	0.7%

hour, which generates lots of high-energy ambient noise in the cafe. These results show that ambient noise is prevalent in our daily life.

2.3.3 Energy Detector

Energy detectors determine the existence or absence of the signal based on a certain criterion. Generally, EDs use the energy in the in-band frequency range as a criterion. The performance of the EDs depends on how precisely they make decisions. Two terms, namely, TP rate and FP rate, are used to express the ED's detection accuracy. The TP rate is calculated as the ratio between the number of detection events categorized as signal and the total number of actual signal detection events. The FP rate is calculated as the ratio between the number of detection events wrongly categorized as signal and the total number of actual noise events. In other words, an ideal ED should achieve 0% FP rate and 100% TP rate. The design of the ED depends on the purpose of the system. If it is important for the system to prevent the effect of noise, the optimization problem is to maximize the TP rate for a given FP rate. On the other hand, if the system considers correct detection of the signal as a top priority, the optimization problem is to minimize the FP rate for a given TP rate.

Generally, EDs save the power consumption by ignoring the noise with energy below a configured threshold. However, if the noise has higher energy than the threshold, EDs classify the noise as a valid signal which corresponds to FP. FP errors caused by high-energy noise induce additional operations to decode valid signals, thus caus-

ing unnecessary power consumption. To eliminate such unnecessary operations, if we design an ED that can detect not only low-energy noise but also high-energy ambient noise in a short time, we can take a step towards more energy-efficient acoustic communication.

2.3.4 FSK Modulation

Along with chirp, an FSK modulation is also widely used in acoustic communication due to its simplicity. There are many applications providing services using FSK modulation-based acoustic communication system [14, 15]. As these FSK acoustic communication services are used in common places such as cafes or department stores, we can assume that people could often encounter the FSK signals in everyday life. As mentioned in Section 2.1, due to limited near-ultra sound frequency range, FSK signals are likely to use nearby frequency range to the in-band frequency range. Since FSK signals can cause more fatal interference than noise, FSK signals should be regarded as a major interference in the ED of chirp-based acoustic communication.

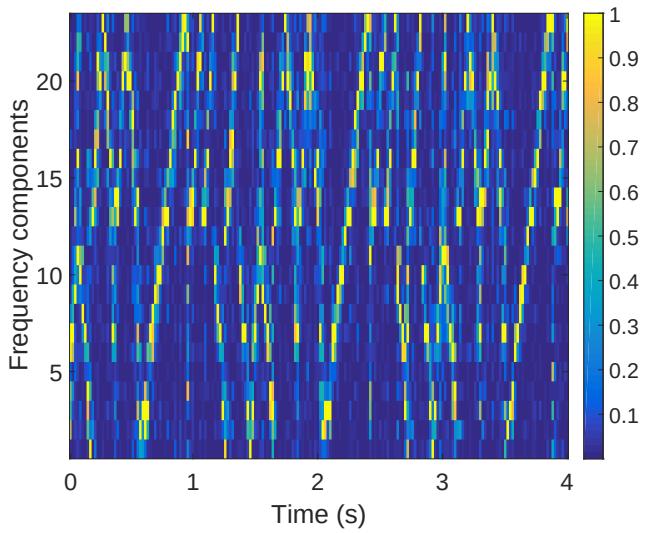
2.4 No Entry: Proposed Energy Detector

In this section, we present the proposed ED, called No Entry. We briefly summarize the entire process of No Entry and then explain each process in detail.

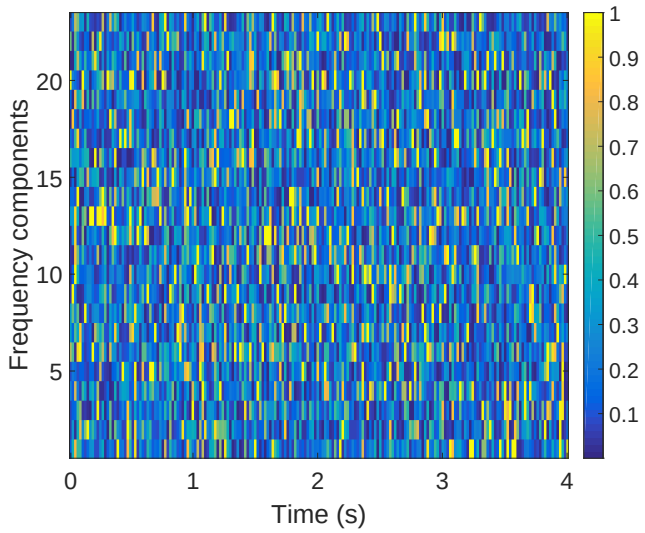
2.4.1 System Overview

When designing an ED, we try to exclude interference that could make a false alarm on the ED. No Entry operates in three processes: 1) low-energy noise filter, 2) ambient noise filter, and 3) FSK signal filter.

In the first process, we observe whether high energy components exist within the in-band. If high energy component is detected in the in-band, the ED assumes the existence of a signal and moves to the next process. However, the presence of high



(a) Normalized PSD of chirp signal



(b) Normalized PSD of ambient noise

Figure 2.4: Normalized PSD plot result of the signal and ambient noise.

energy components in the in-band does not always guarantee the existence of a signal. Ambient noise also has a high energy in the in-band, thus it passes the first process. Fig. 2.4 shows the normalized PSD of a chirp signal and ambient noise. In case of

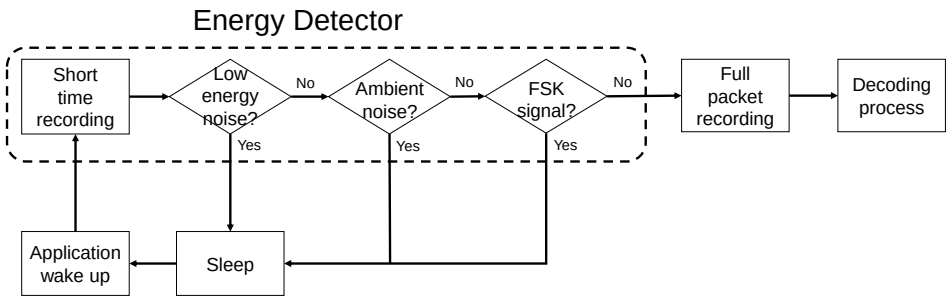


Figure 2.5: Proposed energy detection process flow chart.

a chirp signal, only the frequency components in specific frequency range have high power and the remaining frequency components have low power, which represents W_o and W_v, respectively. On the other hand, the PSD of ambient noise shows disordered patterns. We distinguish between chirp signal and ambient noise by taking advantage of these characteristics.

Even though we deal with the false alarm from the noise, we still have to verify that the received data is a chirp signal. FSK signals are similar to chirp signals from the perspective of time and frequency relationship in that the frequency of FSK signals does not occupy the entire in-band frequency range over a certain time. Therefore, it is difficult to distinguish chirp signals from FSK signals only by checking whether the power is concentrated in a specific frequency range within the in-band. We try to discriminate FSK signals from chirp signals based on whether the frequency is fixed or changed for a certain time, which is the greatest difference between FSK and chirp signals. Throughout these processes, No Entry is able to detect a valid chirp signal. The entire process is shown in Fig. 2.5 and detailed explanation of each process will be described in the next subsections.

2.4.2 Low-Energy Noise Filter

The first process of No Entry is low-energy noise filter which checks whether there exists a high-energy component in the in-band frequency range based on the energy level. The performance of low-energy noise filter is closely related with the detection time (t_{ED}). Applying short t_{ED} may result in lack of time samples, thus leading to the degradation in detection accuracy. On the other hand, to utilize the attributes of a chirp, which are mentioned in Section 2.3.A and will be detailed in the next subsection, t_{ED} should be short enough to divide the in-band into W_o and W_v. Thus, there are two requirements concerning t_{ED} in opposition to each other. We have to use a long t_{ED} to secure the detection accuracy, while the short t_{ED} is desired to observe W_o and W_v. Depending on t_{sym}, the required t_{ED} to meet the second requirement may be too short to securing the detection accuracy. We tackle this challenge by gathering N_{Rx} segments with t_{ED} and making them into a large chunk, which is a similar approach to short-time Fourier transform (STFT) without window overlapping. In this way, we can get a short t_{ED}, which can observe W_v and W_o for each segment, and at the same time, we can get enough time samples to observe in-band energy by combining several segments.

We empirically set t_{ED} to 20 ms and set N_{Rx} to five, which makes a total detection time 100 ms. We denote each segment of the received data in the time domain as $y_i[t]$, and its fast Fourier transform (FFT) as $Y_i[f]$, respectively. Each $Y_i[f]$ has N_{freq} frequency components within in-band frequency range, where N_{freq} is determined by FFT size. For example, if we set bandwidth to 1 kHz, FFT size to 1,024, and sampling rate to 44.1 kHz, the frequency resolution becomes $44,100/1024 \approx 43$ Hz, and hence, N_{freq} becomes 23 (*i.e.,* N_{freq}=$\lfloor$1000 Hz/43 Hz$\rfloor$=23). Let f_k denote the k-th frequency component when the nearest frequency component close to f_{start} is regarded as f_1. The frequency components belonging to the in-band frequency range can be expressed as f_i such that $f_1 \leq f_i \leq f_{N_{freq}}$, thus the in-band energy of the each

segment, denoted as E_i, is calculated as

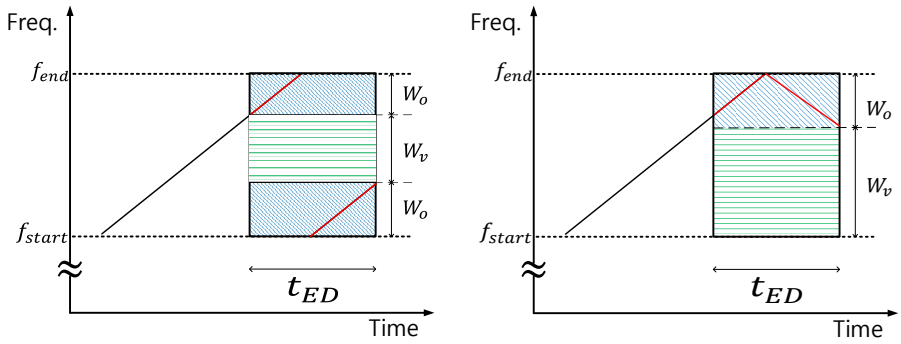
$$E_i = \sum_{k=1}^{N_{freq}} |Y_i[f_k]|^2, \quad i \in \{1, \cdots, N_{Rx}\}. \tag{2.1}$$

We compare the sum of the N_{Rx} segments' in-band energy ($E_{tot} = \sum_{i=1}^{N_{Rx}} E_i$) with a certain threshold (E_{thres}). If E_{tot} is greater than E_{thres}, we go to the next process, assuming that there is a high-energy component in the in-band.

2.4.3 Ambient Noise Filter

To verify that the received data passing the low-energy noise filter is not ambient noise, No Entry examines the PSD of the received data. Since there are too many types of ambient noise, it is impossible to find consistent characteristics. Thus, it would be more reasonable to use the attribute of a chirp signal rather than that of ambient noise as a criterion. We focus on W_o and W_v as a special characteristic of the chirp. The frequency component has a different power depending on whether it belongs to either W_o or W_v. If the frequency components overlap with W_o, then the power of each component is high, otherwise, the power is low. Therefore, the chirp's frequency components are divided into two sets based on the power and location, while those of noise are not clearly divided like chirp signals. We try to find W_o and W_v by gathering the high power frequency components and the low power frequency components, respectively.

As mentioned in the Section 2.3, W_v is either continuous or circularly continuous depending on the detection timing. In addition, W_o could also be either continuous or circularly continuous depending on the symbol combination. Since the ED performs detection process without synchronization, some segments can contain parts of two consecutive symbols. The frequency that the signal sweeps during detection varies depending on the symbol combinations at the boundary of two symbols. Fig. 2.6 shows two different symbol combination examples of the binary chirp. In case of Fig. 2.6a, an up-chirp is followed by a down-chirp. In this case, W_o and W_v are made up of continuous frequency ranges during detection time. In case of Fig. 2.6b, on the other hand,



(a) Continuous

(b) Circularly continuous

Figure 2.6: Different chirp signal combination example.

an up-chirp is followed by another up-chirp. In this case, the first signal's frequency ends at f_{end} and the next signal's frequency starts at f_{start}, which makes W_o circularly continuous. Despite the same detection timing, the different chirp combination makes W_o and W_v different. In other words, we have to consider that both W_o and W_v could be circularly continuous.

We define a metric, named maximum peripheral-to-opposite point peripheral ratio (MOR), to reflect chirp's W_o and W_v. To obtain the MOR, we create two sets of frequency components, S_{max} and S_{op}, for tracking W_o and W_v, respectively. S_{max} and S_{op} are created according to the following procedure.

1. Find $f_{i_{max}}$ having the maximum power among N_{freq} frequency components and add $f_{i_{max}}$ to S_{max}.

2. Set the frequency component $(f_{i_{op}})$ that is $\lfloor N_{freq}/2 \rfloor$ away from $f_{i_{max}}$ to the opposite one, *i.e.*, $\lfloor N_{freq}/2 \rfloor = |i_{max} - i_{op}|$, and add $f_{i_{op}}$ to S_{op}.

3. Compare the power of two frequency components adjacent to S_{max} and add the higher power frequency component to S_{max}. Repeat this process until the number of frequency components in S_{max} reaches N_{max}.

4. Compare the power of two frequency components adjacent to S_{op} and add the lower power frequency component to S_{op}. Repeat this process until the number of frequency components in S_{op} reaches N_{op}.

We expect that S_{max} contains a part of W_o and S_{op} contains a part of W_v. MOR can be defined as follows:

$$\mathrm{MOR}_i = \frac{N_{op} \sum_{k \in S_{max}} |Y_i[f_k]|^2}{N_{max} \sum_{k \in S_{op}} |Y_i[f_k]|^2}, \quad i \in \{1, \cdots, N_{Rx}\}. \tag{2.2}$$

We set the size of the maximum set (N_{max}) and that of the opposite set (N_{op}) differently to reflect different ranges of W_o and W_v. In order to obtain an appropriate MOR through the above procedure, the first selected frequency components, $f_{i_{max}}$ and $f_{i_{op}}$, must belong to W_o and W_v, respectively. We set a condition that restricts t_{ED} to be less than a half of t_{sym}. Accordingly, the maximum W_o is not over BW/2 ($\approx \lfloor N_{freq}/2 \rfloor$) during detection time, which means the range of the absolute value of W_o from point $f_{i_{max}}$ is always less than BW/2. Therefore, we can ensure the selected $f_{i_{op}}$ is always contained in W_v unless $f_{i_{max}}$ is selected out of W_o. Moreover, by specifying the location of the opposite frequency components, noise must have similar PSD to that of chirp not only the ratio between high power range and low power range but also the locations of each high power range and low power range. Thus, specifying the opposite frequency component has an effect of strengthening the effect of the ambient noise filter.

We can get N_{Rx} MORs from the received data and decide whether the received data is a valid signal or noise based on them. However, as shown in the Fig. 2.3, packet has a post-preamble GI. If a segment contains a GI unfortunately, it is difficult to obtain appropriate MOR that correctly reflects the chirp characteristic. In addition, a GI can also affect one or more segments. To minimize the effect of a GI, we select N_{comb} segments out of the N_{Rx} segments based on the in-band energy. Mostly, the in-band energy of the segment including a GI is smaller than that of the segment including a valid signal. We select the top N_{comb} segments with high in-band energy to exclude

the segments including a GI. Therefore, we decide the validity of the received data by comparing N_{comb} MORs with a threshold (MOR_{thres}). Unless all N_{comb} MORs are not larger than MOR_{thres}, we consider the received data ambient noise.

2.4.4 FSK Signal Filter

FSK signals change the frequency in a particular time interval. If we look at the power of an FSK signal in the frequency domain, the power is concentrated at the around of the instantaneous frequency. The relationship between concentrated high power frequency range and the other low power frequency range is similar to the relationship between W_o and W_v of chirp signals. For this reason, neither low-energy noise filter nor ambient noise filter properly discriminates FSK signals.

The biggest difference between chirp signals and FSK signals is whether the frequency is fixed or constantly changed for a certain period of time. Therefore, we try to distinguish between chirp signals and FSK signals by observing whether the frequency constantly changes or not. Because the frequency pattern of chirp signals is so varied depending on the symbol combination, it is difficult to find a method of simply confirming that the received segment is a chirp signal. Instead, we used a method that regards the received segments as an FSK signal if they have a fixed frequency during the detection time. We track the frequency of each segment by observing the maximum power frequency component ($f_{i_{max}}$).

We denote the number of pairs of consecutive segments which have the same maximum power frequency component as Z_{Rx}. However, if we simply track the index of the maximum power frequency component, *i.e.,* i_{max}, there is a problem. Due to the limit of the frequency resolution, the maximum power frequency component can not exactly reflect the instantaneous frequency. Accordingly, despite the same instantaneous frequency, the maximum power frequency component is sometimes changed to the nearby frequency component, which fades our approach to distinguishing the signal through i_{max}.

To compensate this problem, we additionally utilize the second highest power frequency component denoted by $f_{i_{2nd}}$. If i_{max}'s of two consecutive segments are not the same, we check the index of the second highest power frequency component, *i.e.,* i_{2nd}, of each segment. If i_{max} of the former segment equals to i_{2nd} of the latter segment and also i_{2nd} of the former segment equals to i_{max} of the latter segment, this mismatch is considered a transient error and the maximum power frequency components of two segments are considered equal. Finally, we compare Z_{Rx} with a certain threshold (Z_{thres}). When the received data passes all the processes, then No Entry finally determines that there exists a chirp signal.

2.5 Look Inside No Entry

This section covers the issues related with implementing No Entry. The important parameters used in No Entry are analyzed, and then the computational complexity is also discussed.

2.5.1 Parameter Analysis and Discussion

The performance of No Entry depends on various parameters. In this subsection, we will discuss the influence of each parameter. Since the optimization problem of each parameter is a matter of realization, we present a guideline on how to choose the parameters.

Detection time

Detection time (t_{ED}) is the most important parameter in that it gets involved in the entire process of No Entry directly and indirectly. t_{ED} determines not only the detection performance but also the operation time. The number of time samples determines FFT size, which is closely related to the frequency resolution and the computational complexity of No Entry. The frequency resolution is associated with how accurately

we can track W_o and W_v. However, the process of getting MOR is not to find the exact positions of W_o and W_v, but to find some frequency components belonging to the groups. Thus, we do not need an excessively fine-grained frequency resolution. We empirically find that No Entry works well with FFT size greater than or equal to 1,024. The detection time will be further discussed below in conjunction with other parameters.

Set size

Another key parameter is the set size. Two set sizes (N_{max}, N_{op}) play a key role in getting MOR. Getting MOR is the process of tracking W_o and W_v. If we choose inappropriate set sizes, MOR does not reflect the W_o and W_v correctly. In fact, setting the set size is closely related with setting the ratio of W_o and W_v during the detection time, which is closely related with setting t_{ED}. If we shorten t_{ED}, the ratio of W_v within the in-band increases and that of W_o decreases. In this case, we should increase N_{op} and reduce N_{max}. If we lengthen t_{ED}, on the same principle, we should increase N_{max} and reduce N_{op}.

If we take a closer look at the process of getting MOR in (2.2), MOR of the noise is lowered due to a low value belonging to S_{max} or a high value belonging to S_{op}. Since we divide a large value by a small value, increasing the denominator is much influential than decreasing the numerator. This means setting N_{op} to a large value is more effective than setting N_{max} to a small value, which is equivalent to setting a higher ratio of W_v within the in-band. Eventually, getting a effective MOR value comes down to setting t_{ED} to a small value. From this point of view, we can confirm that the necessary condition for detection time is not a strict condition.

2.5.2 Parameter Selection

Detection accuracy varies depending on various parameter values. In the target application, the length of the GI after the preamble is 40 ms. Thus, when we collect five

20 ms-long segments, the number of segments that can contain part of a GI does not exceed three. In other words, at least two 20 ms-long segments do not contain a GI at all. Therefore, we choose $N_{comb} = 2$ to be sure that the selected segments are not affected by a GI, thus minimizing the impact of a GI.

Since W_o and W_v are different for each received data, N_{max} and N_{op} should be set so as to be generally included in W_o and W_v. Since there are infinite kinds of ambient noise, it is impossible to get the optimized parameters for all kinds of ambient noise. However, if we set the parameters that work well over many different kinds of ambient noise, we can expect that No Entry copes with the unexpected noise. First of all, we collect noise data set by generating various types of ambient noise that we can meet frequently in our daily life. The types of ambient noise included in the noise data set are clapping, rattling, sneezing, etc. Then we find a set of the parameters that works best on the noise data set.

Fig. 2.7 shows an example result of Galaxy S7. The graph shows the FP rate by changing N_{max} and N_{op}. We set MOR_{thres} to 99% of the ambient noise data passing the low-energy noise filter. The optimal parameters yielding the lowest FP rate are $N_{max} = 2$ and $N_{op} = 8$, which yield 13% FP rate. The performance does not drastically change when the parameters are selected around the optimal set. However, the FP rate considerably increases if the parameters are selected far from the optimal set.

Based on the result, it can be expected that the performance will not be severely degraded unless the parameters are selected significantly far from the optimal set.

2.5.3 Computational Complexity Analysis

In this subsection, we analyze the computational complexity of No Entry. One of the reasons for using ED is its low computational complexity [10]. Since the computational complexity is closely related to the amount of the signal processing, it is important to operate with as few calculations as possible. Generally, acoustic communication system uses 16-bit pulse-code modulation (PCM) data, and hence the complexity of

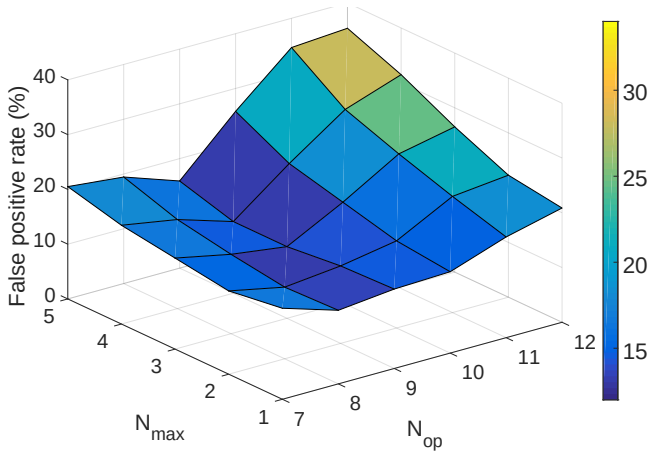


Figure 2.7: The FP rate according to each parameter change.

multiplication dominates the complexity of addition. Thus, we only consider the number of multiplications in the analysis.

We compare the complexity of No Entry with the in-band ED using the same detection time. In-band ED refers to the conventional ED to reject low-energy noise using a single segment unlike the low-energy filter of No Entry which uses multiple segments. In general, the FFT size is chosen to be a power of two. However, since No Entry uses N_{Rx} segments with t_{ED} so that setting the detection time to $N_{Rx}t_{ED}$ can not guarantee that FFT size of the in-band ED is equal to N_{Rx} times of FFT size in No Entry. Therefore, for a reasonable comparison, we assume that FFT size of the in-band ED is equal to N_{Rx} times that of No Entry. In other words, if we denote FFT size of No Entry as N_{FFT}, FFT size of the in-band ED becomes $N_{Rx}N_{\mathrm{FFT}}$. FFT takes part in the largest amount of the computational complexity and the complexity of FFT depends on FFT size. If we denote FFT size as N_{FFT}, FFT computation performs $\frac{1}{2}N_{\mathrm{FFT}}\log_2 N_{\mathrm{FFT}}$ complex multiplications. Secondly, we calculate the complexity of getting the in-band energy. Except for FFT, multiplication and division are performed only in the process of obtaining MOR throughout the process. However, the number of operation is at most N_{Rx}, which is negligible. Therefore, we can compare the computational

complexity taking into account only the complexity involved in FFT computation. As mentioned above, the number of multiplication operations required for FFT computation is $N_{\mathrm{FFT}}(\log_2 N_{\mathrm{FFT}})/2$, so that the complexity of the in-band ED is larger by $N_{Rx}N_{FFT}\log_2 N_{Rx}/2$. Therefore, we conclude that the computational complexity of No Entry is much smaller than that of in-band ED, meaning that No Entry can also maintain the advantage of low computational complexity.

2.6 Performance Evaluation

We evaluate the performance of No Entry in terms of detection accuracy and power consumption. We use Samsung Galaxy S5, S7, and LG G5, G6 as experiment devices.

2.6.1 Detection Accuracy

We compare the detection accuracy of No Entry with three different schemes. We set the in-band ED (Baseline) as a baseline scheme and set J-CS [8] and peak PSD ratio (PPR) [11] as comparison schemes.

Noise filtering

We evaluate the detection accuracy in two different environments with the same signal and noise data sets used in Section 2.3. First, we evaluate the detection accuracy with the noise data set of the offices representing quiet places to verify the ability to handle low-energy noise. Second, we evaluate the detection accuracy with the noise data set of the cafes representing loud places to verify the ability to handle ambient noise. The parameters we use in the experiments are described in Table 2.2. No Entry has several thresholds besides E_{thres}. Thus, we fix E_{thres} to 100% TP rate and observe the accuracy by changing $\mathrm{MOR}_{\mathrm{thres}}$. We present the evaluation results using *receiver operating characteristic* (ROC) curve by changing the threshold defined in each comparison scheme. Fig. 2.8a shows the ROC curves using the noise data set of the offices.

Table 2.2: Experiment Parameter Setting

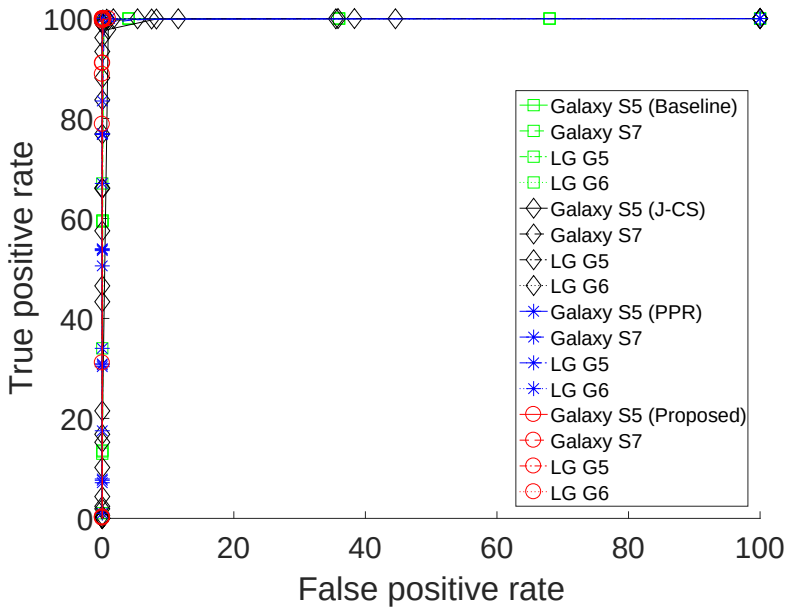
Parameter	Galaxy S5	Galaxy S7	LG G5	LG G6
t_{ED}	20 ms			
N_{Rx}	5			
N_{FFT}	1024			
N_{max}	2	2	1	4
N_{op}	7	8	8	9
N_{comb}	2	2	2	2

As observed in Section 2.3, office environments have little ambient noise. Thus, all of the schemes work properly. Fig. 2.8b shows the ROC curves using the noise data set of the cafes. As expected, the baseline ED shows degraded accuracy due to ambient noise. No Entry and the other comparison schemes, *i.e.,* J-CS and peak PSD ratio, recover the accuracy degradation through their own detection method. As mentioned in Section 2.2, the peak PSD ratio method induces false alarms if ambient noise has a higher energy level in the in-band than that in the out-of-band, which appears in the results of the cafe data. In addition, J-CS, in the case of Galaxy S5, cannot completely distinguish ambient noise and yields degraded performance. On the other hand, No Entry restores the accuracy almost completely as it is designed to withstand ambient noise.

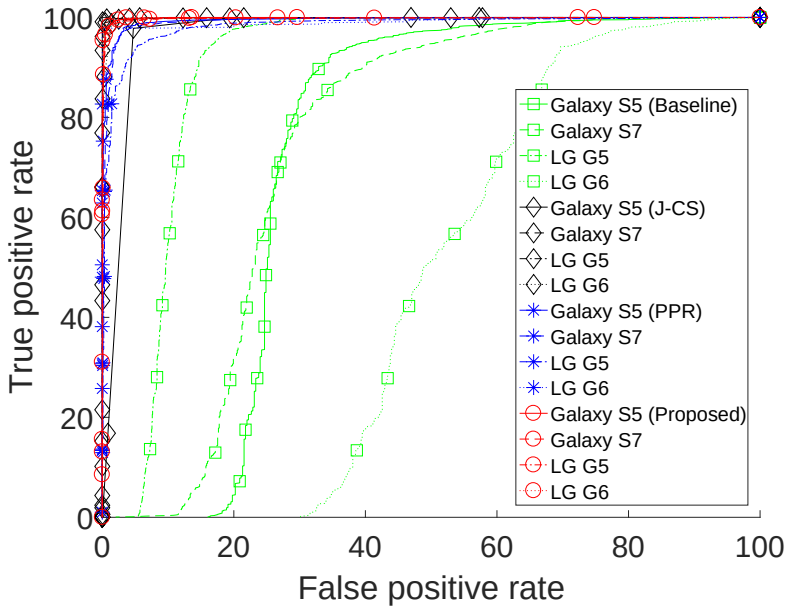
FSK signal filtering

We next evaluate the detection accuracy of No Entry and comparison schemes with an FSK signal. We collect the FSK signal data that is actually being used for order services in cafes [15]. Fig. 2.9 shows the spectrogram of the collected FSK signal. The FSK signal uses a frequency range close to the in-band of the target application, thus possibly causing the false alarms on the ED.

First of all, we verify that No Entry and comparison schemes have difficulty fil-



(a) Office environment



(b) Cafe environment

Figure 2.8: ROC curves in two different noise environments.

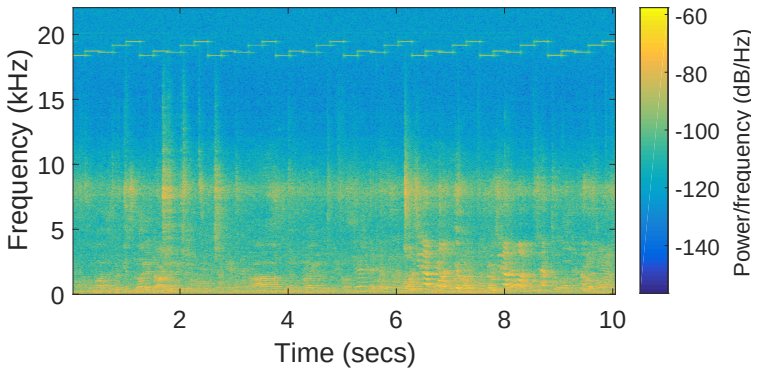
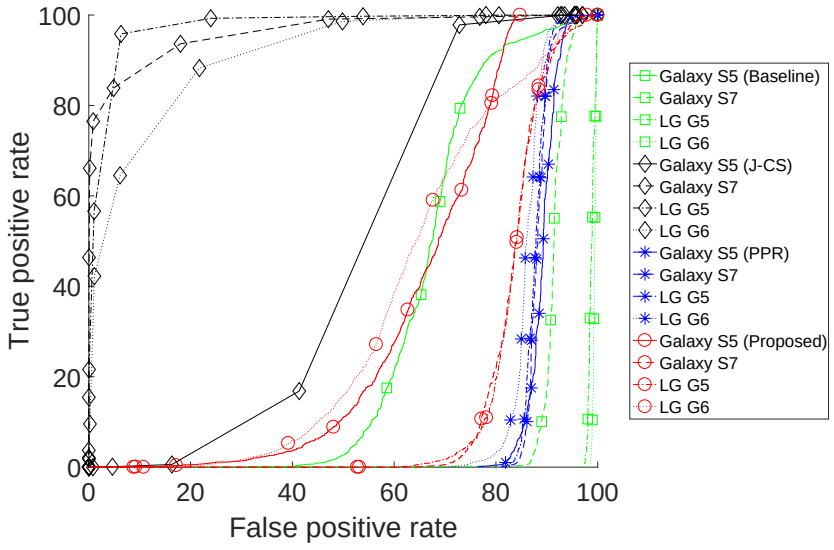


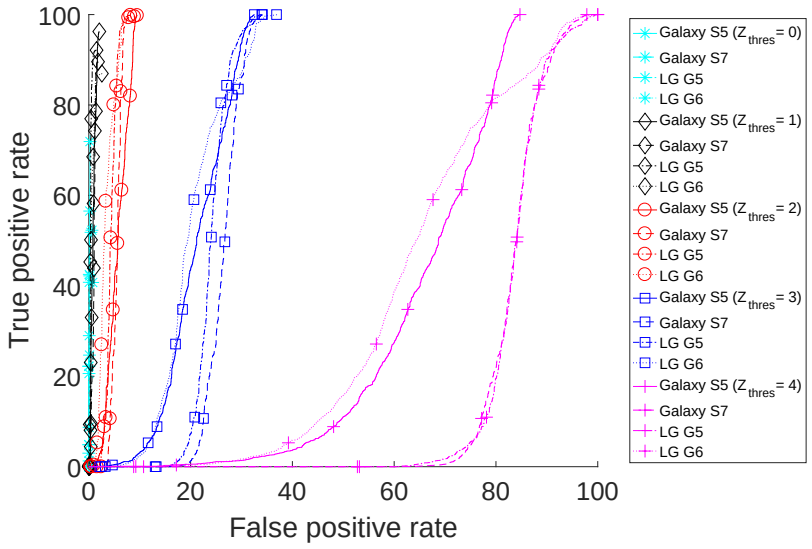
Figure 2.9: Spectrogram of the collected FSK signal.

tering the FSK signal. Fig. 2.10a shows the ROC curves of three comparison schemes and No Entry without FSK signal filter. It is shown that the baseline ED, the peak PSD ratio, and No Entry, which are energy based methods, hardly distinguish the FSK signal. On the other hand, since J-CS detects the chirp signal based on the correlation, it is possible to distinguish the FSK signal to some extent. However, its detection accuracy with the FSK signal is degraded compared with the result with ambient noise, *i.e.*, Fig. 2.8

Fig. 2.10b shows the ROC curves of No Entry with the FSK signal filter. Z_{thres} can range from zero to $N_{Rx} - 1$, so we examine the detection accuracy by changing Z_{thres} from zero to four. Even a valid chirp signal does not always have zero Z_{Rx}, because chirp signal can have a consecutive frequency at the boundary of two symbols. Therefore, When Z_{thres} is set to zero or one, the 100% TP rate is not satisfied. On the other hand, if Z_{thres} is set three or four, the effect of preventing interference is reduced. When we set Z_{thres} as two, No Entry shows the best performance, which outperforms J-CS.



(a) w/o filtering FSK signal process

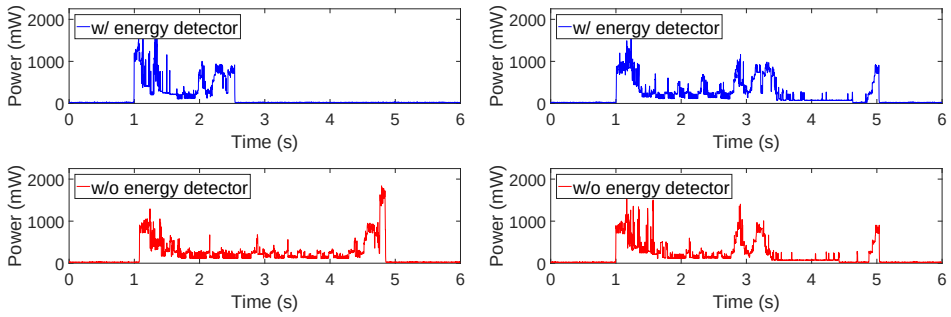


(b) w/ filtering FSK signal process

Figure 2.10: ROC curves with FSK signal data sets.

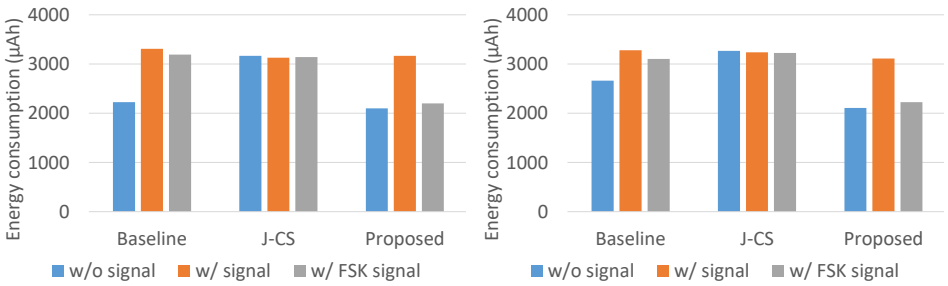
2.6.2 Power Consumption

EDs can save energy by stopping the application after quickly detecting the absence of target signal. We measure power consumption to evaluate how much No Entry can



(a) w/o signal (b) w/ signal

Figure 2.11: Power consumption measurement example.



(a) Office environment (b) Cafe environment

Figure 2.12: Energy consumption measurement result.

save energy. The power consumption of the experiment device is measured using Monsoon power monitor [16]. We use Galaxy S5 for the measurement because the other experiment devices are equipped with built-in batteries, which make the measurement difficult. We implement three types of Android applications, which adopt 1) Baseline ED 2) J-CS, and 3) No Entry, respectively. The application periodically wakes up and tries to receive the signal. Depending on the detection result, the application performs the decoding process if it determines that there exists a signal. If the application determines that there is no signal, it immediately goes to sleep and waits for the next wake-up period.

There are various components consuming power in smart devices. For example,

display and wireless technologies such as Wi-Fi and Bluetooth consume a considerable amount of power. To the extent possible, thus, we shut down the other processes to measure the power consumption of our target application. Furthermore, the authors of [17] find that applications consume substantial power when applications wake up from the idle state. Therefore, we set the wake-up period 10 s and perform the detection process twice at each wake-up to reduce the overhead of the wake-up process.

Fig. 2.11 shows an example of measured power consumption. The results with and without No Entry are similar when the signal exists. On the other hand, if the signal is absent, No Entry detects the absence of the signal and quickly stops working. In this way, No Entry can reduce power consumption of the application. The result using the baseline ED shows similar patterns to this.

We also measure the energy consumption of three types of applications for five minutes in two different places. We perform experiments in three different cases in each place: 1) without any signal, 2) with chirp signal, and 3) with FSK signal. At each place, we place the mobile phone on the table and transmit each signal using a laptop speaker. Fig. 2.12 shows measurement results. For the application with J-CS, it shows consistent energy consumption regardless of the cases because it performs signal detection at the end of the system. The baseline ED reduces energy consumption by filtering low-energy noise. The reduced amount of energy consumption in the cafe without signal is less than that in the office, which shows the limited performance of the baseline ED in the presence of ambient noise. In addition, if the FSK signal exists, the baseline ED hardly saves energy. No Entry, on the other hand, shows outstanding performance in all cases. No Entry works properly in the presence of ambient noise as well as the FSK signal. Compared with J-CS, No Entry reduces the energy consumption by about 30%. Considering the energy consumption caused by wake-up process, the performance of No Entry with the target application itself would be more significant than it appears now.

2.7 Summary

In this chapter, we present No Entry, a novel ED for chirp-based acoustic communication. No Entry avoids both high-energy ambient noise and high-energy interference by utilizing the sweeping characteristic of chirp signals. In particular, we show that ambient noise is prevalent in the everyday life and show that the vulnerability of the conventional ED to ambient noise. We then propose No Entry consisting of three filters, *i.e.*, low-energy noise filter, ambient noise filter, and FSK signal filter. Our evaluation shows No Entry outperforms the comparison schemes from the perspectives of detection accuracy and power consumption. Our future work includes designing an ED, which automatically changes detection period by estimating the severeness of the interference based on the detection result.

Chapter 3

Cameleon: Intelligent Camera Sensor System for Text-spotting Oriented Operation in Mobile Devices

3.1 Introduction

With the recent development of deep learning, many research fields are rapidly developing. Beyond research interests, companies have also adopted deep learning technologies in real life. With the spread and development of mobile devices, there are many efforts to utilize deep learning on mobile devices. In these mobile deep learning systems, the mobile device's sensor data is fed directly to subsequent deep learning models as input. While deep learning model benchmark dataset results promise outstanding performance, in reality, the quality of sensor data has a significant impact on the performance. However, research on sensor control has yet to receive much attention. In order to realize a practical deep learning service on mobile devices, it is necessary to pay attention not only to the development of deep learning technology but also to *generating input sensor data* on mobile devices.

Deep learning models train how to operate through various datasets. Models work well in environments similar to the dataset but suffer from hardness in disparate environments. This phenomenon occurs due to the *unrefined information* in the real world

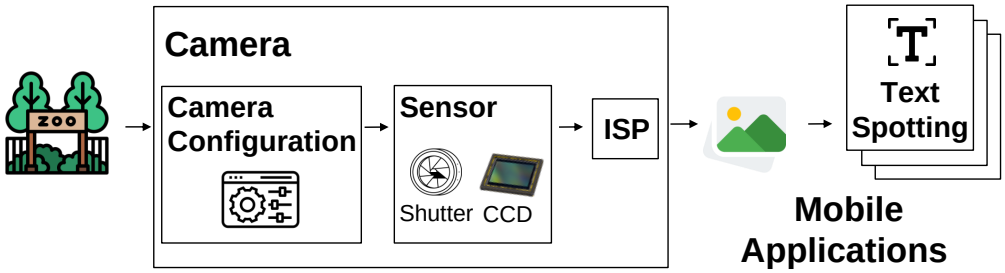


Figure 3.1: Traditional camera pipeline for mobile deep learning application.

compared to the training dataset. Many studies have tried to solve this problem through domain adaptation and data augmentation [18]. These types of research solve the problem temporarily, but they are not fundamental solutions because they require an additional process for the different environments they encounter. Therefore, it will be a fundamental approach if we create input sensor data by manipulating the behavior of sensors to collect information through *information sampling* regardless of the environment.

Mobile devices are equipped with various sensors, such as cameras, microphones, and IMU sensors. Since the camera is the most frequently used sensor in everyday life, vision applications attract considerable attention. We also handle the generating deep learning task-compatible sensor data issue based on the camera sensor. Among many camera-based applications, we focus on mobile text spotting. Mobile text spotting, such as Googlelens and Apple's Livetext, provides services based on the text in the image. Formally, users have to type text themselves to search for or translate text. Mobile text-spotting can change the user's effort to an information-pushing by automatically detecting and recognizing text on the image.

Despite its potential, text-spotting is restrictively available on mobile devices due to the above-mentioned problems. As shown in Figure 3.1, the traditional camera pipeline follows a unified path agnostic of the subsequent mobile applications. Once the input is generated, the ability to process disparate environments specifically for each application is greatly limited. We start off this critical study with a camera control

system focusing on a text-spotting application. To tackle these challenges, we propose a system called Cameleon, **intelligent text-spotting oriented camera sensor control system**. Through our efforts to build a bridge between the text-spotting network and mobile device camera sensor, we find designing Cameleon contains these challenges.

1. **Adaptive and Automatic**: Text spotting networks may behave differently with respect to environmental changes coming from color and brightness changes, even though the target remains the same. The proposed system should work adaptively in these different environments to generate adequate inputs. In addition, the overall process of Cameleon should be automatic. Users should feel natural just like casually taking a picture while operating in the background. This opts for a real-time and efficient operation so that users cannot notice additional operations working behind the scene.

2. **Labeling**: Since selecting camera settings for adaptive environments automatically is a very challenging task, we utilize deep learning to find the adequate setting. In general, deep learning models are trained to produce good results for their own target task. However, in the case of the proposed network, the quality of the output can only be evaluated by the subsequent network's result. In other words, it is difficult to directly generate the labels of the output (i.e., camera setting control) according to the input (i.e., preview image) of the network.

The camera can generate a picture with a short exposure time in bright places. On the other hand, When the light is low, the exposure time should be longer to collect enough light to capture. The high exposure by setting either exposure time longer or ISO high accompanies much noise. In order to reduce the noise, we adopt the method of combining burst shots with a short exposure time as proposed in [19]. However, since the proposed system is not optimized for text-spotting, we have to design a network to control camera exposure.

Determining exposure directly requires broad search spaces, and it is impossible

to make datasets containing all the environments. Therefore, instead of determining camera exposure directly, we build a concept of a *time budget*. The time budget is the total time to generate input data in camera sensors. We categorize the time budgets based on the various experiments. The network determines the appropriate time budget and the number of burst shots within a given time budget. In this way, we can design a practical system operating in various environments.

To handle the second challenge, we use the reinforcement learning method. Similar to the approach [20], we consider Cameleon as an agent, i.e., the output of the Cameleon becomes a policy. The network takes a viewfinder image generated by an auto-exposure as an input. Then the network extracts semantic and scenic information from the preview image through the semantic and illumination networks. We formulate the reward function based on the text spotting results on the image generated by the selected policy (i.e., camera configuration) and the time taken to capture images.

We implement and validate our design through extensive experiments. Compared to the traditional camera pipeline, Cameleon dramatically recovers performance degradation and maximizes the text-spotting model's performance.

The rest of this chapter is organized as follows. We review the related work in Section 3.2. We explain the mobile devices' cameras in Section 3.3, and present the problem of default camera operations as well as our solutions in Section 3.4 and 3.5. We analyze the relationships between camera operations and text-spotting results in Section 3.6. Section 3.7 presents the overall system of the Cameleon and explains how to train the network in Section 3.8. In Section 3.9, we evaluate the performance through extensive experiments. Finally, we conclude the chapter in Section 3.10.

3.2 Related Work

3.2.1 Domain Adaptation

Cameleon is related to solving the domain shift problem. This problem is the main

source of the deep learning model performance gap between the benchmark and real-world environments. To mitigate this problem, many works propose domain adaptation through model fine-tuning and data augmentation [18, 21]. While this approach solves some of the problems, various and time-varying nature of real-world environment mixed with the heavy computation costs of updating the model [22] blocks the fundamental address of the domain shift problem. While some works propose online adaptation, it is still limited to basic applications and the performance is still weak. Cameleon takes a different approach. Instead of dealing with the model itself, it generates the most adequate input for the model by sampling the input information.

3.2.2 Camera Sensor Control for Input Generation

Camera-based vision application is the most prominent sensor-application pair. Existing works focus on modifying the camera pipeline to generate inputs for image classification, semantic segmentation, and object detection [23–25]. While these works focus on optimizing the human perceived quality [26], energy consumption [27], little has been studied about the input generation focused directly on optimizing the model performance.

3.2.3 Text-spotting

To instantiate the proposal, we choose a challenging yet very important vision task, text spotting. Text spotting task is extracting texts from the scene. This can enable machines to 'read' from the scene. Various models are proposed to enhance the text spotting performance [2, 3]. While benchmark performance seems promising, we find that even the state-of-the-art models fail to perform as domain shift occurs. To our knowledge, Cameleon is the first work tackling the input generation for text spotting application.

3.3 Background

3.3.1 Mobile Camera System

Camera controls focus, exposure, and white balance when taking a picture. We focus on the exposure among three components. Camera controls exposures based on ISO, exposure time, and aperture. Since mobile device does not have multiple aperture, we can control ISO and exposure time. ISO refers to camera's sensitivity to light. While the higher ISO produces more brighter image by making the camera sensor sensitive, it inevitably accompanies more noise. Exposure time is the length of time the camera collects light. Longer exposure time brings more light and produces a brighter image, but the image suffers from the noise from hand tremors unless the camera is fixed strictly. Android system provides auto-exposure mode that determines ISO and exposure time based on the statistics depending on the environment.

3.3.2 Camera Capture in Darkness

As the mobile device is usually closely connected to a user, it works in various places and circumstances. Since taking pictures with a mobile device is closely related to brightness, the capture condition varies depending on the time, place, and surroundings. The smartphone camera system controls exposure based on ISO and exposure time as brightness changes. As brightness changes from light to dark, it increases exposure time first. When the exposure time approaches a certain level, the mobile device stops increasing exposure time and increases ISO instead. Since mobile devices have to operate at various levels of motion, the auto-exposure system tolerates noise and raises the ISO to avoid the risk of noise caused by motion when a high exposure time is selected.

3.4 Motivation: Brightness Effect on Text-Spotting

In this section, we present a preliminary study showing the impact of environmental changes on text-spotting models. We select one medicine bottle that contains lots of words and collect several different bright images in the room by controlling the light using a desk lamp. We also collect one additional image in the office under the bright LED lamp for the control group. We use Google Pixel 4 and capture images based on the auto-exposure mode.

Figure 3.2: Impact of environmental (brightness) changes on text-spotting [2] results.

As we can see in Figures 3.2 and 3.3, the text spotting model finds text well in a bright. As the brightness goes lower (left to the right), the text-spotting network suffers from performance degradation, i.e., the number of spotted texts decreases, and this gets worse as it gets darker. TESTR [3] also suffers from the same problem, which indicates that this is not the corner case of the specific model, but the general and essential problem in the text-spotting network. As we can see in the figure, as the room goes darker, the text-spotting performance drops severely. However, if we think of this experiment in a different way, this experiment gives us intuition. Considering that the text-spotting network works well in a light place, it is not only a problem with the text-spotting model but also with generating images. To put it another way, it comes from the underexposure of images. If we can make an image close to that in the light, we can prevent performance degradation. Therefore, we design a system that properly controls camera exposure depending on the environment instead of relying on the auto-exposure operation.

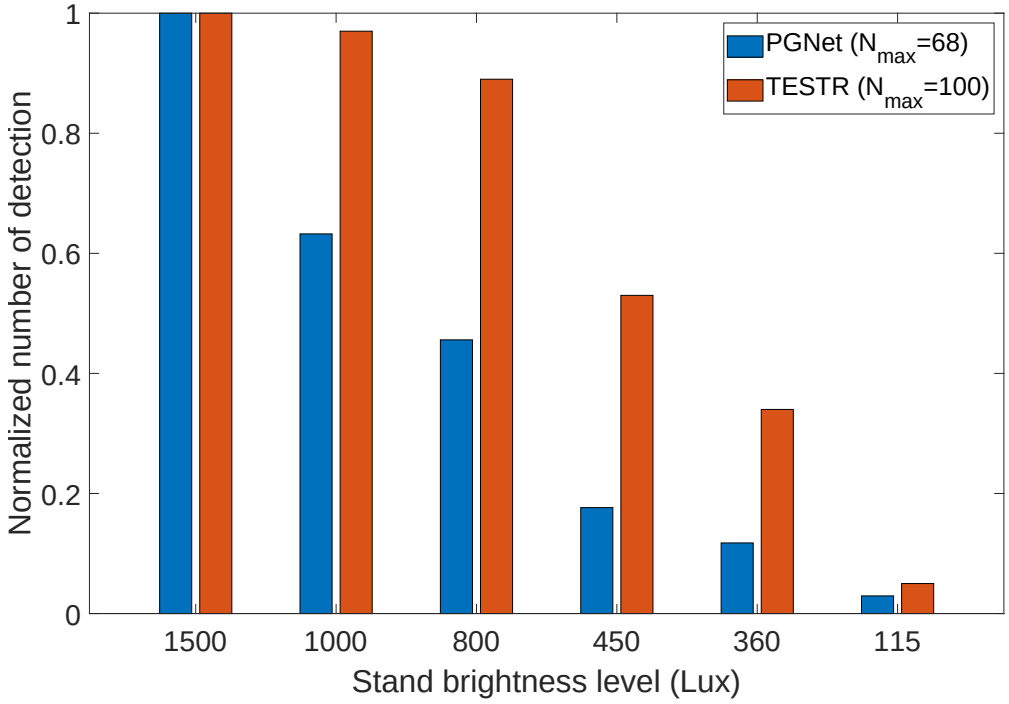


Figure 3.3: State-of-the-art text-spotting networks' performance degradation depending on the environmental changes. [2, 3]

Identifying text in the dark is not a rare occurrence. Most people have experience identifying medicine bottles in a cupboard or looking at the menu at an atmospheric restaurant. If we need text-spotting applications, in that case, it should work in that brightness. Therefore, using mobile device cameras in low light conditions would be common for text-spotting. Table 1 [1] describes situations in daily life in accordance with the brightness (lux). According to [1], smartphone cameras usually start having trouble taking pictures at less than 30 lux, which may be the boundary where the images become noisy. Since the traditional auto-exposure systems accompany lots of noises in dark, it is hard to get text recognizable images.

lux	Description
30,000	sidewalk lit by direct sunlight
10,000	sidewalk on a clear day, but in shadow
1,000	sidewalk on an overcast day
300	typical office lighting
150	desk lighting at home
50	average restaurant
20	restaurant with atmospheric lighting
10	minimum for finding socks that match in drawer
3	sidewalk lit by street lamps
1	limit of reading a newspaper
0.6	sidewalk lit by the full moon
0.3	can't find keys on the floor
0.1	wouldn't walk through the house without a flash

Table 3.1: Description in daily life according to illuminance value [1].

3.5 Challenges and Approaches

In this section, we describe the challenges we face while designing a text-spotting oriented camera control system and our approaches.

3.5.1 Complicated Surrounding Information

To build a system that controls the camera adaptively, the system should be able to figure out the surrounding information correctly. Mobile devices are equipped with a light sensor, for identifying the surrounding brightness. A light sensor measures the surrounding brightness and controls display brightness. However, it is hard to adopt light sensors in camera control for two reasons. First, the target object is usually located backward on the mobile device, on the opposite side of the light sensor. In other words,

the light sensor cannot measure the environment perfectly. The other reason is that the light sensor in the dark does not have enough granularity to divide environmental differences. We change the brightness by controlling the desk lamp and measure the light using the light sensor with three different devices. As we can see in Figure 3.4, the measurement of each device is different despite the same locations. While Pixel 3 and 5 show consistent values, Pixel 4 shows the most reasonable result. However, the measured values are so low that it is hard to discriminate the environment correctly.

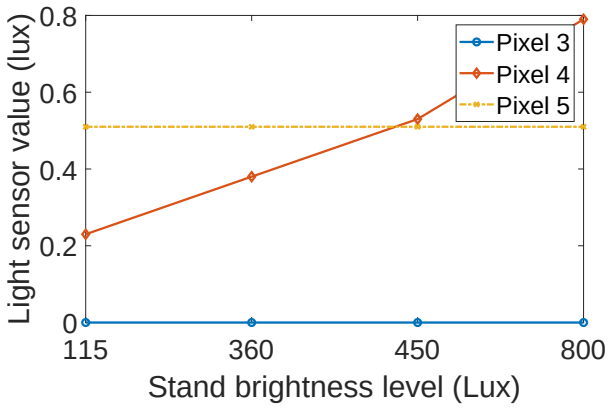


Figure 3.4: Light sensor measurement on different brightness of different devices.

Approach: To analyze the environmental information correctly, we have to observe the surrounding information of the target object. Since the camera scene contains both the target object and environmental information, we use the preview images with a CNN network.

3.5.2 Movement of Mobile Device

If the camera is fixed, it is easy to get a bright image by simply increasing the exposure time long enough even in the dark. However, capturing images with a mobile device being fixed strictly is not usual because most people have hand tremors. Even though we cannot notice the hand's movement sensitively, it significantly affects the capturing process in the dark. If we set the exposure time too long, images get noisy because of

hand movements.

We study the effect of hand movements by collecting images in the dark with two different settings. First, we fasten the smartphone in a holder and increase the exposure time step by step. Secondly, we repeat the same experiment while holding a smartphone in hand instead of a holder. Then, we compare the text-spotting results of each image. As we can see in the figure, when we take a picture with the holder, the number of spotted texts keeps increasing as the exposure time increases. On the other hand, the number of spotted texts for images from hand-held smartphones increases as exposure increases, peaks in the middle, and decreases after the exposure time exceeds 200 ms. In other words, in the dark simply increasing the exposure time is not an adequate solution due to the user's hand movement.

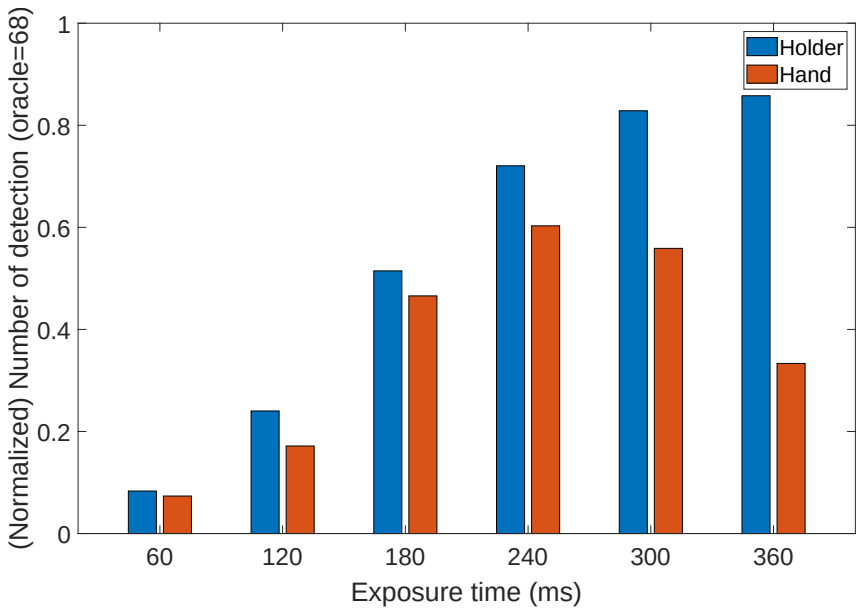


Figure 3.5: Performance variation based on the exposure time. Hand tremor affects performance over a certain exposure time.

Approach: Since simply increasing the brightness is not an adequate solution for the text spotting model, we adopt the burst photography used in [19] instead of single image capture. Burst photography captures burst shots, i.e., multiple images, with low

exposure time and merges them. Merging several images can reduce the Gaussian noise in the image, then the image becomes lighter and less noisy. However, since burst photography is not designed to spot text, the exposure setting is not optimized for text spotting. Therefore, we analyze exposure along with the burst photography method.

3.5.3 Extensive Search Space

For Cameleon, the network output becomes a camera configuration. The camera controls exposure through ISO and exposure time. Both ISO and exposure time have a wide search space. For example, users can manually set ISO and exposure time over the million scales in Android Camera 2 API. Along with camera exposure, we have to control another parameter in burst photography, i.e., the number of burst shots. Putting them together forms three-axis search spaces, making it impossible to select camera configuration directly. We cannot depend on the auto-exposure system because auto-exposure is not designed to work well with burst photography. Therefore, we have to design a new camera control network that determines ISO, exposure time, and the number of burst shots.

Approach: We employ the concept of *time budget*. The time budget is the time elapsed during capturing images, i.e., while taking burst shots. By limiting the time budget, we can narrow down the search space into an acceptable range. In addition, since a system taking too long is not desirable, we can expect the system to be more practical.

3.5.4 Assessment of Capture Settings

Deep learning models are generally trained to produce good results for their target task. Labels are the answers for the task and loss functions are set as the differences between network outputs and labels. Thus, the network is trained to operate to minimize the loss function. However, in the case of Cameleon, the quality of the output can't be evaluated directly, but can only be evaluated with respect to the results of

the subsequent application, here, text-spotting. In other words, it is difficult to directly generate the labels of the task (camera sensor control) according to the input (preview image) of the network. Moreover, it is difficult to train networks through model-based learning because backpropagation is not performed in intermediate modules such as the burst photography module.

Approach: We design an image quality estimator similar to [28]. Quality estimator estimates how much an image fits the text-spotting network. Given the scene, the quality estimator gives high scores to the images with better text-spotting results. We can make labels by estimating the quality of images of various camera control settings.

3.6 Experiment Settings

In this section, we investigate the relationship between camera configurations and text-spotting networks. We control three components (i.e., exposure time, ISO, and the number of burst shots) and analyze the effect of each component to determine the system operations.

3.6.1 Experiment Setting and Metric

We conduct further experiments following Section 3.4. To clarify the effect of camera configurations, we select an object that text-spotting model [3] can spot texts perfectly, i.e., detecting and recognizing all of the texts in the scene perfectly. We use TESTR, provided by the authors, trained using the Total Text [29] dataset. To verify the text-spotting results with both detection and recognition, we use precision, recall, and word level Levenshtein distance [30]. The definitions of precision and recall are as follows

$$Precision = \frac{TP}{FP + TP}, Recall = \frac{TP}{TP + FN}, \tag{3.1}$$

where it becomes 1 when the detection matches ground truth. Levenshtein distance (edit distance) is the difference between two words. In other words, the number of times re-

Metric	Time budget			
	100 ms	250 ms	500 ms	1000 ms
Precision	100	95.0	95.0	100
Recall	42.1	94.7	100	100
Edit distance	-50	-7	-5	0

Table 3.2: Text-spotting results based on the time budget.

quired to insert, delete, and replace while one word matches another. It is zero when the two words are exactly the same.

3.6.2 Time Budget

We first observe the impact of time budget. The time budget is determined by how much time we need to capture images to work well with the text-spotting model. We fix ISO as 1000 and the number of captures as five. As the time budget increases, exposure time also increases.

As we can see in Table 3.2, text-spotting performance dramatically varies based on the time budget. First of all, as we can see in the result of precision, even though it fails to find a text in the image, it does not find text wrongly, i.e., false positive. The precision dropped in 250 and 500 ms does not come from false positives but from detecting two words as one word. However, when it comes to recall, the text-spotting model fails to find texts if the time budget is low, but it finds more texts as the time budget increases. With extensive experiments, we find that the time budget demands up to 1 s. We also verify that the time budget above 1 s does not make a big difference. Because burst photography uses several shots, it also requires a minimum exposure time for each shot. If we increase the time budget too much, the exposure time of each shot also increases, so it suffers from hand movement. Therefore, we set the upper bound of the time budget as 1 s, and we select the final candidates of the time budget as 10, 50, 100, 250, 500, and 1000 ms.

3.6.3 ISO

Secondly, we observe the effect of ISO on text-spotting results. For ISO, the camera system uses values lower than default (100) outdoors on a sunny day (over 1,000 lux in Table 1). In this case, we do not have to use our system because the image is light enough to work with a text-spotting network. Therefore, we specify the target of our system where we have to use it. First, if we set the ISO to 100, it makes the image too dark. On the other hand, if we set the ISO to over 1000, the image starts to have noise. Therefore, we set the boundary of ISO from 200 to 1000. We observe that the ISO does not make big differences in short intervals, we set the intervals to 400.

3.6.4 Burst Shot

Finally, we analyze the number of burst shots given the time budget and ISO. The number of burst shots affects text-spotting results. However, the effect of the number of burst shots is trivial for the short time budget. Given long time budgets, we cannot set a low number of burst shots because it makes the exposure time of each shot long, which raises problems in each image. Based on the previous experiment in Section 4, an image longer than 250 ms suffers from problems due to hand movement. Therefore, we consider 3, and 5 captures in a low time budget under 250 ms and consider 5 and 10 for the time budget exceeding 250 ms.

3.7 System Design

3.7.1 System Overview

The goal of Cameleon is to analyze a given scene in a mobile device camera situation and perform capturing with an optimal camera configuration to achieve appropriate sensing data (image) for text spotting. In order to achieve this purpose, the policy network and burst imaging module were used in this study. We design the system as a

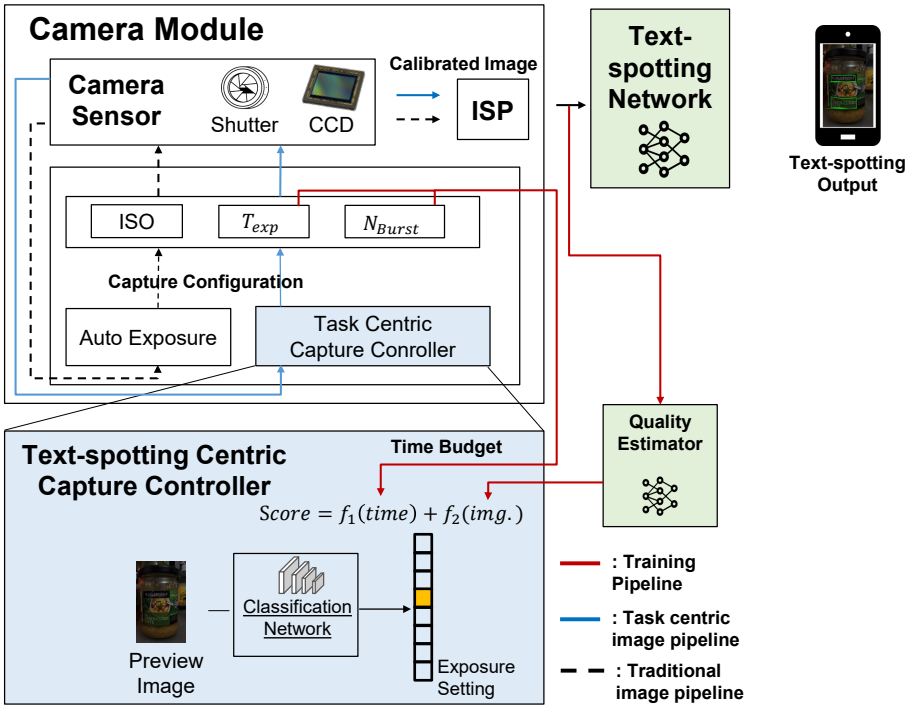


Figure 3.6: System overview of Cameleon.

classification network. The network selects one of the camera configuration options as a classification output based on the scene's information. After the camera captures the scene based on the selected camera configuration, raw sensor data goes into the burst imaging module.

3.7.2 Classification Network

The network takes a viewfinder image, a conventional auto-exposure-based preview image, as input and generates outputs capturing parameters. We design the network as a classification network and adopt the mobileNetV2 [31] as a backbone network. The network analyzes semantic features and chooses an appropriate options as an output.

3.7.3 Burst Imaging Module

The burst imaging method is used for signal processing of raw input captured based on the camera setting as the output of the Policy Network. The method used in this study is a module of [19], and a simple corresponding model was used to achieve the optimal capturing process by focusing on the camera sensor control, the purpose of this study. The image derived through the module is used for classification network. As a follow-up study, it is inferred that higher performance and efficiency can be aimed by jointly connecting the differentiable ISP module and the capturing module and learning it.

3.7.4 Quality Estimator

The quality estimator estimates the quality of the images for text-spotting. We adopt inceptionV3 [32] as a backbone network and add classification header behind the network. To train the quality estimator, we first train the network using COCO-text dataset and total-text dataset [29, 33]. We make labels based on the text-spotting results. We use F-score (F_1) on each scene for text detection results and use the Levenshtein (edit) distance [30] for recognition results.

$$\mathrm{Score} = \mathrm{w}_1 \times \mathrm{F}_1 + \mathrm{w}_2 \times \mathrm{D}_{\mathrm{edit}},$$

, where w_1 and w_2 are the weights of each score. To train the quality estimator for our task, we collect 4,000 images with various camera configurations under the various environments (objects, places, light, etc.). We make text labels using total-text label tools. After making scores based on the text spotting network, we conduct fine-tuning using our datasets. As we can see in the Table 3.3, quality estimator shows high accuracies, Pearson, and Spearman correlations. Now, we can make labels based on the quality estimator instead of making hand-crafted labels.

	Accuracy	Pearson	Spearman
Top-1	71.6%	0.947	0.948
Top-2	91.6%	0.938	0.938
Top-3	96.7%	0.914	0.910

Table 3.3: Quality estimator performance.

3.8 Training Network

3.8.1 Data Collection

We collect various image captures using different capture configurations for each scene using our custom Android application. We control the time budget to six different levels. Within each time budget, we control the number of image bursts to control the exposure time for each scene. By controlling the exposure time, we focus on finding the sweet spot between the trade-off between brightness and hand tremor. Then for each image capture, we control the ISO and the sensor's sensitivity to three different levels.

3.8.2 Label Distribution Learning

We can generate labels by collecting images with various exposure settings and scores them. If we make the label in a hard-labeling (i.e., one-hot encoded form), which selects only one capture setting as a label, it can be a problem. We adopt the label distribution learning method [34] to alleviate the above-mentioned problem. The label distribution learning observes the relationship between labels and composes labels as a distribution form. As shown in Figure 3.7, we can make labels into distribution using a soft-max function. We use the KL divergence as a loss function.

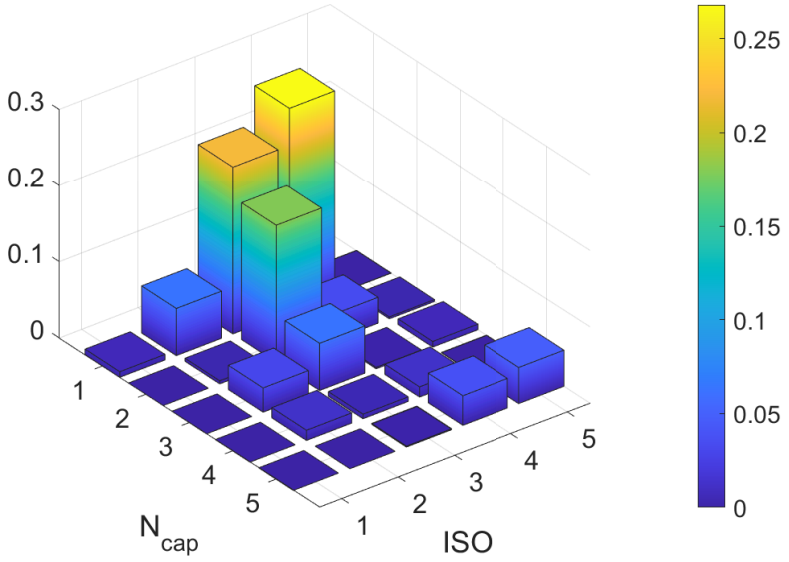


Figure 3.7: Distribution form label example.

3.9 Performance Evaluation

In this section, we evaluate the performance of Cameleon. We implement Cameleon as an Android application. We train the system in the edge server equipped with GeForce RTX 2080 GPU. Since it is hard to get the original burst photography code, we use the publicly published code in Python instead [35]. We implement Cameleon using Android Camera 2 API, which provides burst shots functions and exposure settings for each shot. We collect images by varying the brightness through stand lighting in a room. We make the test dataset by collecting another object not used for the training dataset by varying the brightness in a different place.

3.9.1 Top-5 Accuracy

We evaluate the performance using two different metrics. First, we evaluate the top-5 accuracy to verify the validity of the proposed unsupervised training method and classification model. While the classification model trained with PGNet label shows

the top-5 accuracy of 80%, in the case of TESTR, it shows a slightly smaller accuracy of 65%.

As we can see in the results, the accuracy is not much high. We generate labels based on text spotting results, which means the results have high similarities in certain environments. If the results of each exposure value are similar overall, most probability values show similar results, and in this case, it is often difficult to select the correct result. If we select an exposure with a value slightly lower than the highest result, the accuracy becomes lower, but it is not a big problem for overall operation. Therefore, we need to evaluate the performance by observing the end-to-end improvements.

3.9.2 End-to-end Evaluation

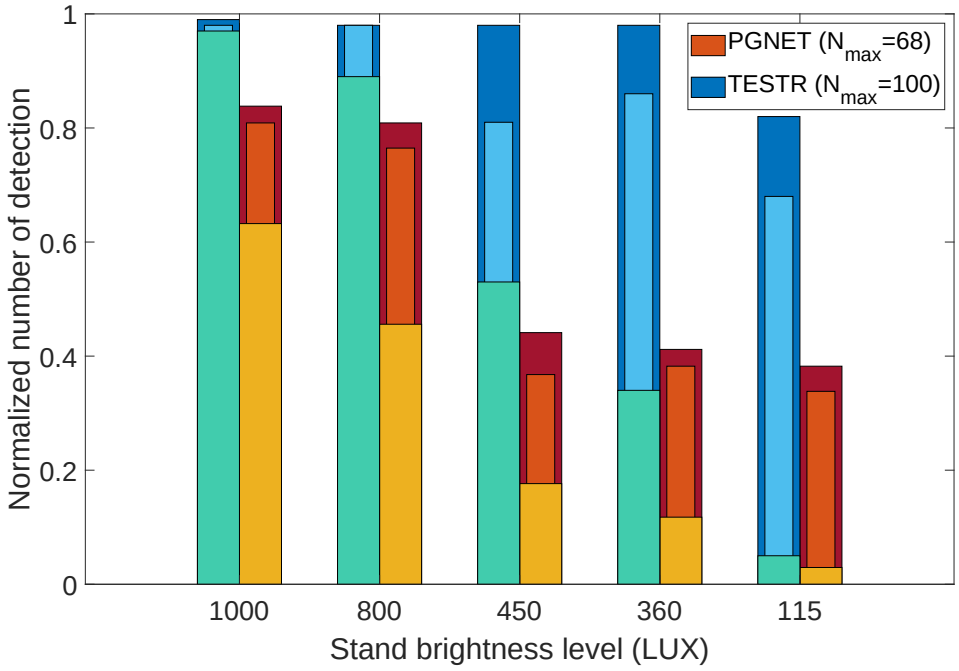


Figure 3.8: Improvement results of Cameleon.

Next, we evaluate the end-to-end improvements, i.e., how many texts the text-spotting network finds through Cameleon compared to default camera operation. There

Model	Brightness (Lux)				
	1000	800	450	360	115
PGNet	27%	68 %	108 %	225%	1050%
TESTR	1%	10%	53%	153%	1260%

Table 3.4: Improvement gain for each model.

are three bars in the graph. The lowest bar means the baseline (i.e., default camera) performance, and the largest bar means the ideal improvement gains. The bar in the middle means the selected exposure settings using our classification models. As we can see in Figure 3.8, the performance of Cameleon shows dramatic improvements in both models. The degree of improvement differs for each model, but both models find more texts than that in the baseline.

As we can see in the graph, the two models show different improvement gains. Not only does TESTR perform better than PGNet in default, but it also shows much higher improvement gains in Cameleon. TESTR and PGNet show a similar number of detection results in the dark, but the gap in the number of detection between the two models becomes even more comprehensive in the light. It means that TESTR is more robust than that of PGNet to environmental changes. We can find similar patterns in improvement gain results. TESTR shows more significant improvement gains even in the darkest places. A slight improvement in input image induces a much more significant improvement in text-spotting results. We also can verify that the degree of improvement is not directly correlated with classification accuracy.

3.10 Summary

We present Cameleon, a text-spotting-oriented intelligent camera control system. We point out the conventional camera module has limitations to use in mobile text-spotting, because it does not operate in a way to fit text-spotting. We also attack the this problem

brings about underperformance in text-spotting network. We solve this problem by designing camera control system to fit text-spotting network. We design overall system and propose a training method. We verify that the problem of conventional camera operaions and show that Cameleon outperforms default camera. Our future work includes online network training methods and optimized operations in mobile devices.

Chapter 4

Concluding Remarks

4.1 Research Contributions

In this dissertation, we have addressed the systems that improve the efficiency and performance of new types of various sensors-based applications in smartphones.

In Chapter 2, we present No Entry, a new ED for chirp-based acoustic communication. No Entry utilizes the sweeping properties of the chirp signal to avoid both high-energy ambient noise and high-energy interference. In particular, it shows that ambient noise is prevalent in everyday life and that the existing ED is vulnerable to ambient noise. Then we propose No Entry consisting of three filters: a low-energy noise filter, an ambient noise filter, and a FSK signal filter. We implement prototype Android application and measure the power consumption using Monsoon power monitor. Detection accuracy of No Entry shows that true positive rate is more than 90% when false positive rate is 1% even with severe interference. No Entry also reduces energy consumption by about 30% compared with the state-of-the-art scheme.

In Chapter 3, we present Cameleon, a text-spotting-oriented intelligent camera control system. We point out the conventional camera operation is not optimized for mobile text-spotting, because it does not operate in a way to fit text-spotting. We solve this problem by designing an text-spotting-oriented camera control system. We verify

that the problem of conventional camera operaions through extensive experiments and show that Cameleon outperforms default camera.

4.2 Future Research Directions

As further improvement on the results of this dissertation, there are two research items regarding camera control system.

First, we plan to devise online training method to update model continuously. Since camera does not capture all of the camera settings, it is challenging to train the network.

Second, we plan to optimize the network for mobile operations. Since mobile device lacks in computational capabilities and batteries, we will make the network efficient.

Bibliography

[1] Google, “Night sight: Seeing in the dark on pixel phones,” 2018. [Online]. Available: https://ai.googleblog.com/2018/11/night-sight-seeing-in-dark-on-pixel.html

[2] P. Wang, C. Zhang, F. Qi, S. Liu, X. Zhang, P. Lyu, J. Han, J. Liu, E. Ding, and G. Shi, “Pgnet: Real-time arbitrarily-shaped text spotting with point gathering network,” *AAAI. AAAI*, pp. 2782–2790, 2021.

[3] X. Zhang, Y. Su, S. Tripathi, and Z. Tu, “Text spotting transformers,” *arXiv preprint arXiv:2204.01918*, 2022.

[4] W. Mao, J. He, and L. Qiu, “CAT: high-precision acoustic motion tracking,” in *Proc. ACM MobiCom*, 2016.

[5] B. Zhou, M. Elbadry, R. Gao, and F. Ye, “BatMapper: Acoustic Sensing Based Indoor Floor Plan Construction Using Smartphones,” in *Proc. ACM MobiSys*, 2017.

[6] R. Nandakumar, K. K. Chintalapudi, V. Padmanabhan, and R. Venkatesan, “Dhwani: secure peer-to-peer acoustic NFC,” in *Proc. ACM SIGCOMM*, 2013.

[7] H. Lee, “Aerial acoustic communication using chirp signal,” Ph.D. dissertation, Seoul National Univ., 2014.

[8] S. Ka, T. H. Kim, J. Y. Ha, S. H. Lim, S. C. Shin, J. W. Choi, C. Kwak, and S. Choi, "Near-ultrasound communication for TV's 2nd screen services," in *Proc. ACM MobiCom*, 2016.

[9] Q. Wang, K. Ren, M. Zhou, T. Lei, D. Koutsonikolas, and L. Su, "Messages behind the sound: real-time hidden acoustic signal capture with smartphones," in *Proc. ACM MobiCom*, 2016.

[10] S. Atapattu, C. Tellambura, and H. Jiang, *Energy detection for spectrum sensing in cognitive radio*. Springer, 2014.

[11] H. Lee, T. H. Kim, J. W. Choi, and S. Choi, "Chirp signal-based aerial acoustic communication for smart devices," in *Proc. IEEE INFOCOM*, 2015.

[12] Z. Sun, A. Purohit, R. Bose, and P. Zhang, "Spartacus: spatially-aware interaction for mobile devices through energy-efficient audio sensing," in *Proc. ACM MobiSys*, 2013.

[13] Loudness Comparison Chart. [Online]. Available: http://www.sengpielaudio.com/TableOfSoundPressureLevels.htm.

[14] LISNR. [Online]. Available: http://lisnr.com/.

[15] YAP. [Online]. Available: https://yap.net/en/.

[16] Monsoon power monitor. [Online]. Available: https://www.msoon.com/LabEquipment/PowerMonitor/.

[17] S. Park, D. Kim, and H. Cha, "Reducing energy consumption of alarm-induced wake-ups on android smartphones," in *Proc. ACM HotMobile*, 2015.

[18] J. Yang, S. Shi, Z. Wang, H. Li, and X. Qi, "St3d: Self-training for unsupervised domain adaptation on 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 10 368–10 378.

[19] S. W. Hasinoff, D. Sharlet, R. Geiss, A. Adams, J. T. Barron, F. Kainz, J. Chen, and M. Levoy, "Burst photography for high dynamic range and low-light imaging on mobile cameras," *ACM Transactions on Graphics (ToG)*, vol. 35, no. 6, pp. 1–12, 2016.

[20] Z. Wang, J. Zhang, M. Lin, J. Wang, P. Luo, and J. Ren, "Learning a reinforced agent for flexible exposure bracketing selection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1820–1828.

[21] S. Yang, Y. Wang, J. van de Weijer, L. Herranz, and S. Jui, "Generalized source-free domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 8978–8987.

[22] R. Bhardwaj, Z. Xia, G. Ananthanarayanan, J. Jiang, Y. Shu, N. Karianakis, K. Hsieh, P. Bahl, and I. Stoica, "Ekya: Continuous learning of video analytics models on edge compute servers," in *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*, 2022, pp. 119–135.

[23] J. Guo, H. Gu, and M. Potkonjak, "Efficient image sensor subsampling for dnn-based image classification," in *Proceedings of the International Symposium on Low Power Electronics and Design*, ser. ISLPED '18. New York, NY, USA: Association for Computing Machinery, 2018. [Online]. Available: https://doi.org/10.1145/3218603.3218618

[24] C.-T. Wu, L. F. Isikdogan, S. Rao, B. Nayak, T. Gerasimow, A. Sutic, L. Ainkedem, and G. Michael, "Visionisp: Repurposing the image signal processor for computer vision applications," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 4624–4628.

[25] E. Onzon, F. Mannan, and F. Heide, "Neural auto-exposure for high-dynamic range object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[26] S. Diamond, V. Sitzmann, F. Julca-Aguilar, S. Boyd, G. Wetzstein, and F. Heide, "Dirty pixels: Towards end-to-end image processing and perception," *ACM Transactions on Graphics (SIGGRAPH)*, 2021.

[27] M. Buckler, S. Jayasuriya, and A. Sampson, "Reconfiguring the imaging pipeline for computer vision," 10 2017, pp. 975–984.

[28] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE transactions on image processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

[29] C. K. Ch'ng, C. S. Chan, and C. Liu, "Total-text: Towards orientation robustness in scene text detection," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 23, pp. 31–52, 2020.

[30] V. I. Levenshtein *et al.*, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, no. 8. Soviet Union, 1966, pp. 707–710.

[31] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[33] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, "Coco-text: Dataset and benchmark for text detection and recognition in natural images," *arXiv preprint arXiv:1601.07140*, 2016.

[34] X. Geng, “Label distribution learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1734–1748, 2016.

[35] A. Monod, J. Delon, and T. Veit, “An Analysis and Implementation of the HDR+ Burst Denoising Method,” *Image Processing On Line*, vol. 11, pp. 142–169, 2021, https://doi.org/10.5201/ipol.2021.336.

초 록

스마트폰의 보급과 더불어 다양한 종류의 센서를 장착한 모바일 기기가 늘어나고 있습니다. 하드웨어의 발전과 다양한 종류의 센서는 모바일 기기에서 새로운 많은 것을 가능하게 하였습니다. 최근 몇 년 동안 많은 연구자들은 다양한 센서 기반의 어플리케이션을 제안하고 있습니다. 예를 들어, 스마트 기기의 마이크와 스피커를 이용한 대기중 음파 통신은 최근 많은 관심을 받고 있습니다. 또한, 카메라 센서 기반의 글자 감지 기술 (Mobile Text-Spotting) 역시 학계와 산업계에서 활발한 연구 주제 중 하나입니다. 하지만, 이러한 새로운 기술을 실생활에서 사용하기 위해서는 효율성과 실용성 두 가지 측면에서 접근이 필요합니다. 본 논문에서는 모바일 센서 기반의 스마트폰 어플리케이션의 효율적인 동작을 위한 두 가지 시스템을 제안합니다.

첫째로, No Entry는 처프 기반 음향 통신 시스템을 위한 새로운 에너지 검출기입니다. No Entry는 처프 (Chirp) 신호의 주파수가 변하는 특성을 활용하여 높은 에너지의 실생활 노이즈 뿐만 아니라 다른 변조 기법의 음향 신호또한 감지합니다. 검출 정확도와 전력 소비를 평가하기 위해 Android 프로토타입 어플리케이션을 구현하였고, 최근 제안된 다른 방법들과 비교하여, 제안하는 에너지 검출기는 에너지 소비를 30% 줄이면서도 더 높은 검출 성능을 보여주었습니다.

두번째로, 모바일 기기에서의 딥러닝 어플리케이션을 위한 카메라 센서 컨트롤 시스템을 제안합니다. 딥러닝 모델은 많은 발전을 이루어 왔지만, 데이터셋을 기반으로 동작하는만큼 데이터셋과 상이한 환경에서 동작할 경우 성능이 저하되는 문제가 있습니다. 이를 해결하기 위해 글자 감지 모델의 최적화된 동작을 위한 지능형

카메라 센서 제어 시스템을 설계합니다. 사람 눈에 좋은 이미지를 생성하는 일반적인 카메라 작동과 달리 지능형 카메라 센서 제어 시스템은 환경에 따라 글자 감지 모델에 최적화된 형태로 카메라 센서를 컨트롤합니다. 전체적인 네트워크의 설계와 더불어 학습 방법, 데이터 수집 방법을 제안하였습니다. 또한, 광범위한 실험을 통해 문제가 있음을 확인하였고, 제안하는 시스템이 다양한 환경에서 잘 동작하는것을 확인하였습니다.

본 논문에서는 스마트폰 센서를 활용한 어플리케이션의 효율적인 동작을 위한 시스템을 제안하였습니다. 상용 스마트폰에서 두 가지 시스템을 구현하였고 광범위한 실험을 통해 성능을 검증하였습니다.

주요어: 스마트폰, 모바일 센서, 모바일 어플리케이션, 모바일 딥러닝
학번: 2015-20885