An Adaptive and Integrated Multimodal Sensing and Processing Framework for Long Range Moving Object Detection and Classification

by

Tao Wang

Thesis Committee:

Professor Zhigang Zhu (Advisor), CUNY City College
Professor Ioannis Stamos, CUNY Hunter College
Professor YingLi Tian, CUNY City College
Dr. Ajay Divakaran, SRI-Sarnoff

A dissertation submitted to the Graduate Faculty in Computer Science in partial fulfillment for the requirements of degree of Doctor of Philosophy,
The City University of New York

2012

Abstract

In applications such as surveillance, inspection and traffic monitoring, long-range detection and classification of targets (vehicles, humans, etc.) is a highly desired feature for a sensing system. A single modality will no longer provide the required performance due to the challenges in detection and classification with low resolutions, noisy sensor signals, and various environmental factors due to large sensing distances. Multimodal sensing and processing, on the other hand, can provide complementary information from heterogeneous sensor modalities, such as audio, visual and range sensors. However, there is a lack of effective sensing mechanisms and systematic approaches for sensing and processing using multimodalities. In this thesis, we described a systematical framework for Adaptive and Integrated Multimodal Sensing and Processing (thereafter, the AIM-SP framework) that integrates novel multimodal long-range sensors, adaptive feature selection and learning-based object detection and classification for achieving the goal of adaptive and integrated multimodal sensing and processing. Based on the AIM-SP framework, we have made three unique contributions. First, we have designed a novel multimodal sensor system called Vision-Aided Automated Vibrometry (VAAV), which is capable of automatically obtaining visual, range and acoustic signatures for moving object detection at a large distance. Second, multimodal data, acquired from multiple sensing sources, are integrated and represented in a Multimodal Temporal Panorama (MTP) for easy alignment and fast labeling. Accuracy of target detection can be improved using multimodalities. Further, a visual reconstruction method is developed to remove occlusions, motion blurs and perspective distortions of moving vehicles. With various types of features extracted on aligned multimodal samples, we made our third contribution on feature modality selection using two approaches. The first approach uses multi-branch sequential-based feature searching (MBSF) and the second one uses boosting-based feature learning (BBFL).

To my wife Yun and my parents for their love and support

Acknowledgments

First and foremost I would like to express my sincerest gratitude to my thesis advisor Prof. Zhigang Zhu, who has greatly supported and guided me throughout years of my PhD research with his patience and knowledge. This thesis would not have been completed without his commitment and diligent efforts. I am also grateful to Prof. Zhu who always encourages me to make my work into publications (Appendix E) as well as brought my attention to numerous R&D opportunities. I can never forget the support from him which helps me build upon strong academic and industrial experience so that I can secure a job position quickly, even before I graduate. He is a source of inspiration. What I learned from him is far more than I expected and will benefit me in the rest of my life.

I would also like to thank my other thesis committee members, Prof. Ioannis Stamos, Prof. Yingli Tian, and Dr. Ajay Divakaran, for their time commitments, selfless assistance, invaluable feedback and great patience at all stages of this PhD process. Thanks to Prof. Stamos who taught me the fundamental concepts in computer vision when I was in the entry level of my PhD study. With his influence, I started to love to do the research in vision-related fields. I am also grateful to Prof. Tian who brought insightful ideas in my thesis proposal and helped me work toward right directions. I also appreciate Dr. Divakaran for his willingness to use his expertise in signal processing with down-to-the-detail comments he provided for not only this thesis but also other published work with him.

Moreover, I am also indebted to researchers at Wright Patterson Air Force Research Laboratory during my summer internships. My special thanks go to Dr. Clark N. Taylor for his thoughtful discussions. I would also like to thank others at AFRL, including Dr. Kevin Priddy, Dr. Erik Blasch, and Ms Olga Mendoza-Schrock for their supports. I would also like to express my gratitude to Pastor James Chun-min Yeh, other brothers and sisters (especially Sister Meng), at Dayton Chinese Christian Church (DCCC), for their love and care during my ten-week stay at Dayton, Ohio in 2011. Their kind words and actions will never be forgotten. Mostly, I'd like to thank our loving God who is taking a large place in my life.

I would like to thank all friends and colleagues at the City College Vision Computer Lab to make a friendly working environment and share their interesting ideas. Thanks go to Edgardo Molina, Hao Tang (now a professor at BMCC), Wai Khoo and other newly joined PhD students. I also want to thank two visiting scholars during my thesis years, Dr. Yufu Qu and Dr. Rui Li for their helpful collaboration and inspiring discussions.

This work has been supported by the Air Force Office of Scientific Research (AFOSR) under Award #FA9550-08-1-0199 and the 2011 Air Force Summer Faculty Fellow Program (SFFP), by the National Collegiate Inventors and Innovators Alliance (NCIIA) under an E-TEAM grant (No. 6629–09), and by a PSC-CUNY Research Award. The work is also partially supported by National Science Foundation (NSF) under Award #EFRI-1137172 and Award #CNS-0551598, and Army Research Office (ARO) under Award #W911NF-08-1-0531.

These supports have greatly inspired and facilitated the research reported in my PhD thesis.

I cannot end here without thanking my parents, for their greatest supports and absolute confidence in me, without which I would not have survived the PhD process. My final words go to Yun, my dear wife, who gave me two lovely kids during my PhD years. That gives me a lot of momentum as well as huge amount of pressure, but that is also the major reason that I feel life is enjoyable and worth living.

Table of Contents

Chapter 1

1 Introduction

Recently, research and development efforts in moving object detection and classification are gradually shifting their emphases from only analyzing visual information to using multiple sensing modalities. Remote object signature detection is becoming increasingly important in non-cooperative and hostile environments for many applications (Dedeoglu, et al, 2008; Li, et al, 2008). These include: (1) remote and wide area surveillance in frontier defense, maritime affairs, law enforcement, and so on; (2) perimeter protection for important locations and facilities such as forest, oil fields, railways and high voltage towers; (3) search and rescue in natural and man-made disasters such as earthquakes, flooding, hurricanes and terrorism attacks. In these situations, target signature detection, particularly signatures of humans, vehicles and other targets or events, at a large distance, is critical in order to watch out for the trespassers or events before taking appropriate actions, or make quick decisions to rescue the victims, with minimum risks. Although imaging and video technologies (including visible and IR) have had great advancement in object signature detection at a large distance, there are still many limitations when they are used in non-cooperative and hostile environments due to intentional camouflage and natural occlusions. Audio information, another important data source for target detection, still cannot match the range and signal qualities provided by video technologies for long-range sensing, particularly under a variety of large background noises. For obtaining better performance of human tracking in a near mediate range, Beal, et al.

(2003) and also Zou and Bhanu (2005) have reported the integrations of visual and acoustic sensors. By integration, each modality may compensate for the weaknesses of the other.

Multimodal sensing and processing have become very active research topics. The *multimodal sensing* part deals with multiple sensory modalities thus involves sensor coordination, data synchronization, and data integration. On the other hand, the *multimodal processing* part deals with multimodal features thus need data processing, feature extraction and classification all for multiple modalities. However, in most research works, these two components are handled rather separately; sensor developers mainly care about building multimodal sensory systems for specific tasks and processing scientists overwhelmingly focus on well prepared multimodal data for their researches. There is a connection gap between two groups and a lack of a close-loop evaluation and feedback between each other. As a particular example such as long range moving object detection using both audio and video, we have to not only design a sensing system to multimodal data acquisition, but also develop a method to synchronize and process those data collected from multiple sensor sources. And decisions made based on those data or features can help us evaluate the necessity of different sensor sources. Therefore, there is a strong need of a systematic and underlying framework that connects all steps. For this reason, we have developed a unified *Adaptive and Integrated Multimodal Sensing and Processing* (AIM-SP) framework that integrated smart data collection, adaptive feature selection and optimal object detection and classification for achieving the objective of

adaptive and robust multimodal sensing for situational awareness, in particular, in vehicle and human detection.

In the rest of this chapter, we will give a general description of our goal in multimodal sensing and processing, and three main objectives in the framework in Section 1.1. Then Section 1.2 discusses challenges to achieve the goal. Section 1.3 presents an overview or our approach. Section 1.4 summarizes our three unique contributions. Section 1.5 shows an outline of the structure of the dissertation.

1.1 Goals

The ultimate goal of our research is to apply AIM-SP framework to a wide range of different tasks (human detection, vehicle detection, bridge monitoring, anomaly detection, and etc.) using the same inference framework through optimal feature selection and classification ensemble learning. For illustrating the effectiveness of the approach, we will particularly focus on multimodal long range moving object (vehicle, human) detection and classification throughout the rest of the chapters. The target applications could be surveillance, traffic monitoring and inspection. Many sensor technologies, such as video, audio, radar, infrared, and ultrasonic could also be used for those applications. However, we will mainly focus on two typical sensor components, audio and video, since they are commonly used in surveillance applications to acquire different target signatures and provide complementary information to each other. So one of our objectives is to align and label samples from audio and video data with separate sensors. Then, multimodal information, such as visual appearance, motion, range, and

acoustic signatures for the same objects could be extracted. Definitely more information can help us make a better decision but there also exists redundant, unimportant or even unrelated information for a specific task. For example, if we only want to distinguish vehicles of different shapes, visual features may dominate the decision, and audio is irrelevant to this task. However, if we want to measure the volumes of the engine sound of particular vehicles, such as a truck vs. a minivan, audio information may influence our decision more. It is very important to select the most representative data or features given a specified task. Although our experiments will focus on specific tasks, we do not want to limit our goal only to those. Therefore, the other objective is to show that sensor selection, feature learning and decision making, even though all related to particular tasks, but a general approach can be developed. In order to achieve the goals, many challenges are involved and discussed hereunder.

1.2 Challenges

Multimodal adaptive sensing systems with various sensor modalities including visual, range and acoustic measurements have found applications in biometrics (Chen, et al, 2010), activity recognition (Petsatodis, 2009), and large area surveillance (Cristani, et al, 2007; Dedeoglu, et al, 2008). However, long range moving object detection using multimodal sensing systems opens up some challenge researching issues. First, are the sensing systems re-configurable and adaptive? Second, are the data collected from the multimodal sensors automatically done or easy to process to generate object signatures? Third, are all features extracted equally important to make a decision for a specific task?

Furthermore, can the system learn from previous results to provide feedback for better feature selection and even adaptive sensor control without redesign the whole system, when giving a new task?

For the first issue, the state-of-the-art sensor technology is not ready for adaptive sensing. In the past, most surveillance systems use cameras only. Now, in additional to conventional visual based sensing systems, there are some systems using other sensors, such as sonar, infrared cameras, or laser Doppler vibrometer (LDV) to detect vehicles on road (Samadi, et al, 2008; Iwasaki, 2008; Wang, et al., 2011; Qu, et al, 2010). It has been shown that the use of multimodal sensors provides better performances in object detection and classification. For acoustic signature acquisition, a microphone or microphone arrays are used in multimodal sensing systems. Those types of sensors need to be placed close to the targets of interest. Many of them need to be fixed at pre-determined locations if an object is in motion and needed to be tracked. Parabolic microphones, which can capture voice signals at fairly large distance in the direction pointed by the microphone, could be used for remote hearing and surveillance. But it is very sensitive to noise caused by the wind or the sensor motion, and all the signals on the way are captured. In the City College Visual Computing Laboratory, we have found another emerging sensing technology for long range audio acquisition that using a laser Doppler vibrometer (LDV) (Zhu, 2004). The LDV sensors were initially designed for industrial and architectural inspection applications. However, they are found to be able to detect acoustic signals at a large distance through the detection of the vibration of the

surface of an object near a sounding target. Therefore, they can be used to perform long-range multimodal surveillance and monitoring by integrating visible and infrared videos.

This new sensing technology leads to the second issue – multimodal data collection and integration. It is a complicated procedure and needs to be considered carefully. Data collected from multiple sensory components are always noisy and not aligned in a way that can be easily processed during multimodal feature extraction and integration. For example, for the same target, the target detection from an audio may not lie on the same time frame as that from a video. This is always the case that we sometimes hear the sound of a coming vehicle before we actually see it. For vehicles' visual detection, most methods (Gupte, et al., 2002; Hsu et al., 2006) assume that the desired vehicles can be detected by image differencing. Then various kinds of vehicle features like shape, texture, etc. are extracted easily to make the vehicle classification straightforward. However, several environmental variations will significantly affect the accuracy of vehicle classification. This will be even more the case for long-range vehicle detection and inspection, where the sensors (cameras) can only be set in a remote location. In such a scenario, the standpoints of view from the sensors to a road could be constrained due to large distances, and occlusions such as trees and other facilities. This will result in the failure of vehicle detection and degrade the accuracy of later vehicle classification and recognition. Another environmental variation is that the perspective views (ranges, directions) of captured vehicles which also vary greatly. When a vehicle is observed along a lane, it will have different appearances/resolutions in different video frames over the period of time. Also, the video data of the vehicle could be in a low resolution and subject to motion blur.

Vehicles' visual images should be reconstructed from multiple video images to solve those issues. The vehicles' sounds do not face the occlusion and perspective view issues from the video, and can also provide complementary information, such as loudness and sharpness, for distinguishing different types of vehicles. An LDV can capture the acoustic signatures of a moving target, such as humans and vehicles, at a large distance; however, its signal strength is affected by the vibration and the reflection of a background surface. Thus, audio enhancement should be performed for obtaining better acoustic signatures.

After multimodal alignment, visual reconstruction and audio enhancement, various types of features can be extracted. However, not all features from multimodal sensors are equally important to make a decision for a given task. Some can provide complementary information, but some will provide redundant or even irrelevant information. Also, the fusion of features that are obtained from different modalities usually results into a large feature vector. Many feature vector reduction techniques (Wall et al., 2003; Guironnet et al., 2005; Chetty & Wagner, 2006) are applied to reduce the data from higher dimensional space into lower dimensional space. However, those techniques are mainly based on the data themselves. We will explore feature modality selection as an alternative approach to reduce the size of features meanwhile improving the performance in multimodal feature integration. It refers to choose different types of modalities from which could be different sensor sources or heterogeneous features extracted from a single sensor source. The question is, from an available multimodal feature set, which modalities should be selected to accomplish a specified task? The utility of those modalities could be changed for different tasks. As the optimal feature subset changes over time, how confidence the

feature modality selected with which the task is accomplished, is an open problem for multimodal feature fusion and classification.

We note that in most multimodal sensing research, sensor modalities, control feedback, and target signatures are mostly defined by the designers, and mostly use a model-driven approach. On the other hand, in machine learning research, data classification and event decision are only based on the pre-defined datasets, which are mostly data-driven. There is little work that addresses the adaptive data collection and feature selection based on the system performance. Therefore, there is a strong need for interaction between the model-driven approach and the data-driven approach in order to solve the last issue – to have an adaptive and integrated multimodal sensing-processing control system.

1.3 Overview of Our Approach

To respond all challenges described above, we first introduce an Adaptive and Integrated Multimodal Sensing and Processing (AIM-SP) framework shown in Figure 1.1. The AIM-SP framework provides the flexibility to put particular tasks and sensing data in a unified framework in order to provide feedbacks for collecting the most useful signals from a set of heterogeneous sensors. The scenes and targets determine what the effective multimodal sensors are for data acquisition, where the users and tasks specify what kinds of techniques need to be used to achieve the goal. There are two

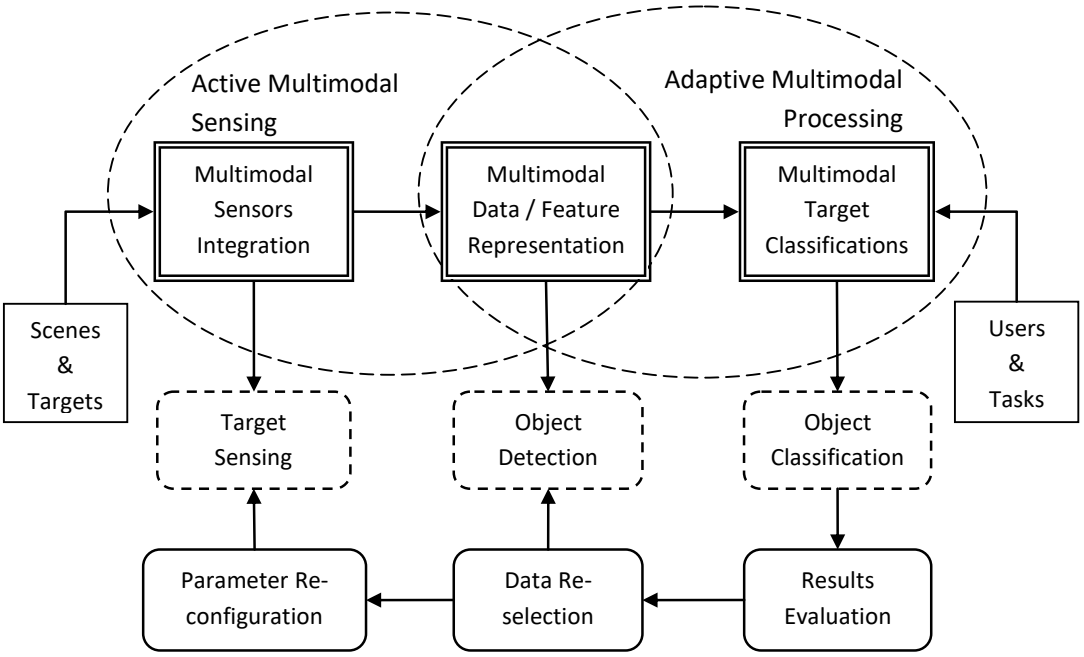


Figure 1.1 Multimodal sensing and processing framework

interrelated parts, active multimodal sensing and adaptive multimodal processing. Three phases are processed across theses two parts: *multimodal sensors integration* in the sensing part, *multimodal target classification* in the processing part, and *multimodal data representation and feature extraction*, which builds a connection between the data collected from the sensors and the data used for multimodal target classification. There is also an evaluation chain corresponding to the three phases, which provides a close-loop feedback. Each phase serves some particular objectives. For example, the sensor integration phase is used for active target sensing; then data are collected and represented for better object detection using multimodalities. Because the synchronization at the data representation level, features can be easily aligned and correlated to each other for improving classification. Then feature modality selection can

evaluate those representative features and sensor modalities. The goal of this work is to provide a close-loop interaction between sensing and processing, therefore benefiting both sensor design engineers and algorithm development scientists. As an example, we will target on long range moving object detection and classification, particular, vehicles and humans. At each phase, we will also address our novelties and contributions.

In the *multimodal sensory integration* phase, we design an adaptive multimodal sensing system based on a novel concept of Vision-Aided Automated Vibrometry (VAAV), which provides visual assistance for the long range remote hearing using an active vibration sensor – LDV, in fast and automated target selection via intelligent focus and distance measurement. Moving object detection at a large distance is very challenging in surveillance and inspection. The VAAV system has unique features for automatic or interactive reflective surface detection and laser pointing through the visual assistance for better target detection. The main contribution of this work is the collaborative operation of a dual-PTZ-camera system and a laser pointing system for the long-range acoustic detection. To our knowledge, this is the first piece of work that uses a PTZ stereo for automating the long-range laser-based voice detection. Meanwhile, the combination is a natural extension of the already widely used PTZ-camera-based video surveillance system towards multimodal surveillance with audio, video and range information.

The *multimodal data representation* phase builds a connection between the sensors integration and the features extraction. For moving vehicle detection and classification, data collected from multiple sensors are represented in a Multimodal Temporal

Panorama (MTP) that aligns three main modalities: shape, motion and audio, in the same time axis so that multimodal features can be extracted synchronously. The MTP facilitates the synchronization and integration of the information across the three modalities, both for automatic and interactive vehicle and traffic analysis, thus providing more succinct and reliable information for tasks like moving vehicle detection and classification using visual, motion, and acoustic information. In the MTP representation, we have two objectives in mind: providing mechanisms for automatic vehicle detection and a user-friendly GUI for training data labeling for later vehicle classification. For these purposes, while generating the MTP, real-time processing and detection are also implemented. With the help of the MTP, we also develop a visual reconstruction algorithm for moving vehicles, which can remove occlusion, motion blur and perspective distortions. The purpose of reconstruction is to provide an automated and cleaned data for easy labeling and improved classification. Meanwhile, the reconstructed results and the corresponding original images shot are stored and displayed for comparison and archival. The spectrum of the vehicle's sound is also displayed to enhance the detection of the vehicle using acoustic information. Note the time spans of the vehicle in video and audio may not be the same since we usually hear the sound of the vehicle before actually see it. Therefore the MTP provides a very effective user interface to visualize and analyze the alignment of the video and acoustic information of passing-by vehicles, thus facilitating the joint detection and classification of vehicles using both visual and audio information.

After multimodal data alignment and synchronization, the system enters the *multimodal target classification* phase. Multimodal features are extracted. We analyze various types

of visual features and audio features, and select those features that could provide optimal classification performance for a given task. We first provide a flexible multi-branching feature searching technique that is based on sequential forward selection algorithm. It selects a number of good features and their branches at different levels of combinations. We notice that the feature extraction and selection are task-dependent. Given different tasks, the same features may play different roles. In this work, we design two different types of classification tasks using the same set of features on the same dataset and provide a thorough study on the feature selection and combinations for vehicle classification. For robust feature selection, we also propose a boosting-based feature learning technique to select a number of same or different feature modalities at each weak learner to further improve the classification accuracy.

We believe that the proposed architecture along with the new techniques will be a valuable addition to both industrial and scientific research, and may also open new research areas in encouraging the cooperation between sensor engineers and processing scientists.

1.4 Summary of Contributions

Based on the unified Adaptive and Integrated Multimodal Sensing and Processing (AIM-SP) framework (Fig. 1.1), we have made three unique contributions:

1. A novel Vision-Aided Automated Vibrometry (VAAV) multimodal sensor system is designed that is capable of obtaining visual, ranging and acoustic signatures for moving object detection at a large distance. The integrated system greatly

increases the performance of the LDV remote hearing and therefore increases its feasibility for audio-visual surveillance and long range object inspection and detection applications.

2. A Multimodal Temporal Panorama (MTP) approach is developed for multimodal data representation and alignment, which facilitates target detection and data labeling. It provides multimodal information including visual appearances, motion signatures and acoustic information. In addition, it provides the capability to reconstruct vehicles' visual appearances so that motion blurs, occlusions and perspective distortions can be removed. It also provides a very effective user interface for training data labeling in both video and audio domains.
3. A multi-branch feature searching (MBFS) algorithm and a boosting based feature learning (BBFL) algorithm are proposed to select the representative feature modalities. The effectiveness of multimodal feature selection and combinations are thoroughly studied through empirical studies. The performance between MBFS and BBFL is also compared based on our own dataset.

We also provide an audio-visual vehicle (AVV) dataset for long range moving vehicle detection and classification. It also contains moving people with different activities. The data are collected at two locations, one at a two-way local road and the other at a multi-lane highway, using the multimodal sensing system we designed. However, we mainly use the data of moving vehicles at the local road in support of our feature modality selection experiments.

1.5 Outline of the Dissertation

This chapter introduced the problem of multimodal sensing and processing for moving object detection and classification, and presented the overall goals of the thesis. It also outlined our approach to the problem and summarized our main contributions. The remaining chapters are organized as follows.

- Chapter **Error! Reference source not found.** describes the state of art in multimodal sensing and processing, particularly in surveillance applications.
- Chapter 3 describes our multimodal sensor design (VAAV) and multimodal sensor integration. It first describes the sensory components and their calibration, and then introduces a control framework for active control and adaptive sensing.
- Chapter 4 presents the dataset we built using the VAAV system, then describes the MTP-based multimodal data. Reconstruction algorithm based on the MTP is also described with error analysis.
- Chapter 5 describes the feature extraction for each modality and a technique for acoustic feature enhancement.
- Chapter 6 presents the multi-branch feature searching and boosting based feature learning techniques for the feature modality selection and multimodal classification.
- Chapter 0 summarizes our work and provides a discussion of the advantages and limitations of the work. It also provides some suggested directions for future research in this field.

Chapter 2

2 Related Work: a Literature Review

Surveillance is important for the security and wellness of societies in both civilian and military domains. In the past, surveillance systems typically have single modalities, and the majority of them used visual sensors. In the last few years, multimodal surveillance systems are attracting more and more attention. Video surveillance is one of the fastest growing sectors in transforming from single sensor modalities into multimodalities, for example tracking moving object using both color and infrared cameras (Torresan et al., 2004; Davis et al., 2005; Conaire et al., 2006). However, these systems are still not capable of detecting suspicious events that cannot be "seen", such as screaming, gunshot and etc. Therefore, Audio-Visual (AV) systems have been used in some degrees for surveillance tasks, for example, recognizing human fighting (Dedeoglu et al., 2008), and understanding human activities (Cristani et al., 2007). A few of these systems (e.g. Zhu et al., 2007) have applied multimodal sensing techniques to audio and visual surveillance at large distance.

An automated, multimodal surveillance system, which uses the information from multiple sensor modalities, attempts to automatically detect and track objects of interest, and also models and analyzes activities of usual and/or unusual behaviors for those interested objects. Such a system should consist of a set of adequate sensors, reliable methods in the acquisition of the sensing data, and effective integrating algorithms. All parts of the system are equally important and need to be integrated to produce reliable output for

detection, tracking and recognition. Several multimodal surveillance systems and related techniques have been discussed in a recent edited book by Zhu and Huang (2007).

Using multi-sensory data or multi-information is not limited to multimodal surveillance systems; it is a topic of great interest in other multimodal systems, such as biometrics, multimedia, multimodal medical imaging, and remote sensing. Here, we will only briefly discuss the relations between multi-biometrics, multimedia systems and multimodal systems, which are closely related to multimodal surveillance. Sometimes multi-biometrics, multimodal and multimedia systems are used interchangeably, for example, in Atrey et al. (2006). However, they have some important differences. *Multi-biometric* systems are those which utilize, or are capable of utilizing, more than one physiological or behavioral characteristics of humans (such as ear, iris, face, gesture, and voice) for enrollment, verification or identification (Sanderson and Paliwal, 2004, Zou and Bhanu, 2005; Thieme, 2007). *Multimodal* systems interpret and regenerate (by fusing) information presented from different inputs (sensory data) to make a decision. These sensor data could be in the forms of not only human signatures (biometrics), but other information such as vehicle signatures, scene description and other context information (Zhu and Huang, 2007). Multimodal systems support users multiple ways of responses according to their preference and needs. *Multimedia* systems, on the other hand, refer to a user's adaptation of a system's perceptual capability, and are more concerned about issues of human computer interaction (HCI). An example is presented in Atrey et al. (2006). In other words, multimedia systems focus more on control and integration of output information. A lot of recent work on multimedia content analysis can be found in

a book c edited by Divakaran (2009). In term of interaction, multimodal systems provide the ability that allows users receive multimodal input and are able to respond by using those modalities. In this survey, we will focus more on *multimodal* systems for *surveillance* applications, mostly using vehicles' signatures.

The rest of review is mainly divided into three parts: sensing modalities, multimodal surveillance systems, and multimodal data fusion methods. The sensing part will briefly describe several commonly used sensor modalities and their capabilities for the surveillance. These are mainly discussed in Section 2.1. Section 2.2 describes the three processing steps: low-level feature extraction, intermediate-level data processing, and high-level classification and recognition in two commonly used modalities: video and audio. With these two kinds of surveillance systems, we will give some details of low-level feature extraction and intermediate-level data processing, followed by a discussion of the need of multimodalities. Section 2.3 focuses on the multimodal data fusion. Section 2.4 presents motivations behind our approach and summarizes its relation to prior work.

2.1 Sensor Modalities

Sensors are important since they are the front ends of a surveillance system. The performance improvement using multimodal sensing has direct impact on the detection accuracy and false alarm rate of each sensor, operating in a complex and cluttered environment. Therefore, understanding the sensing characteristics of each sensor modality is a critical step. Since our major interest is surveillance at a large distance and in a wide area, those sensor modalities applicable to long-range scenarios will receive

more attention here. Figure 2.1 shows a few examples of commonly used sensors in market. Table 2.1 lists a few important parameters of various sensor modalities for use in long-range, large-area surveillance: measured entities (light, sound, etc.), sensing principles, field of view (FOV), image resolution/ or measurement accuracy, sampling rate, sensing range (distance), sensor size and cost. We evaluated the parameters of those sensors presented here, and assigned their capabilities from low, medium to high by relative comparison.

Figure 2.1 Various sensor modalities: a few examples

Table 2.1 Comparison of sensors and their important parameters

Sensor	Regular EO	Omni camera	PTZ	Thermal/IR	LDV	Mic-array	Lidar	Radar	Sonar
Measure	light	Full view	ROI	Temperature	Vibration/ acoustic	Sound	Range	Objects and ranges	Objects and ranges
Principles	Light to electricity	Imaging + optics	Imaging + locomotion	Thermal radiation	Doppler interference	Sound waves	TOF: laser	TOF:audio microwave	TOF: sound
FOV	Normal array	360º	Controllable	Normal array	Point to area	Omni-directional	Point to area	Point to area	Point to area
Resolution/ Accuracy	Medium to high	Low to medium	Medium to high	Low to medium	High accuracy	medium	High res &accuracy	Medium	Low
Sampling Rate	60 Hz	60 Hz	60 Hz	60 Hz	11, 22.5, 45 KHz	>16 KHz	>33 KHz	>192 KHz	41-96 KHz
Range	Medium (100m) to far (1Km)	Small to medium	Medium (100m) to far (1Km)	Medium (100m) to far (1Km)	Medium (100m) to far (1Km)	Small	Medium (100m) to far (1Km)	Medium (100m) to far (1Km)	Medium (100m) to far (1Km)
Size	Medium to small	Medium to small	Medium to small	Medium to large	Medium to large	small	Medium to large	large	Small to large
Cost	low	medium	Medium to low	Medium to high	high	low	high	high	medium

2.1.1 Electro-optical (EO) sensors

Electro-optical (EO) sensors may be the most commonly used sensors for surveillance applications. For the application of tracking a moving target at a distance, a PTZ sensor, particularly with high zoom ability, is used to provide the control sufficient to focus on interesting targets. However, one PTZ camera may fail in the tracking task when the tracked object moves too fast, or is occluded. Therefore the use of multiple visual sensors to track moving objects of interest is commonly used in a wide-area surveillance application. Typically, a panoramic (omnidirectional) sensor is used as a master sensor along with one or more PTZ sensors as slaves to track multiple objects (Cui et al., 1998; Scotti et al., 2005; Yao et al., 2006). As an assumption of some systems, all the participating sensors are precisely geo-calibrated in order to accurately localize objects. The requirement of pre-calibration is not a big issue if all the sensors are stationary or only undertake pan/tilt/zoom operations; but this will be a challenging problem if some

of the sensors are mounted on moving platforms, either aerial or ground (Xiao et al., 2008). One of the major factors affecting the quality of the image data collected from EO sensors is the *illumination* of the environment. And most optical sensors are limited to well-lit conditions only, whether the lighting is provided by the Sun during the day or by artificial and planned lighting of dark areas during night and/or day.

2.1.2 Thermal or Infrared (IR) Sensors

Thermal or infrared (IR) imaging uses the heat radiation or the infrared spectrum that is independent of the ambient light in the area to be imaged. It is useful for night-time and hidden-area surveillance. Infrared imaging technology has been used by the military and civilian systems for surveillance. Now it has become a mature main-stream technology. There is a large body of literature on using infrared sensors in surveillance applications. Crebolder et al. (2003) presented a technical report describing the general role of infrared sensors in large military reconnaissance systems. In surveillance and biometrics applications, thermal sensors have been employed to fuse thermal data with optical data in order to detect and classify features on human faces (Heo et al., 2004; Kong et al., 2005), etc..

2.1.3 Laser Range and Vibration Sensors

Recently, laser sensing technologies have created a lot of opportunities in surveillance systems. There are two types of laser sensors that have mostly applied in surveillance application, the laser rangefinders and the laser Doppler vibrometers (LDVs). Both of them follow the principle of sending a laser beam towards the object and measuring the

reflected beam. Many laser range sensors actually have an imaging camera in addition to the laser itself, which is basically used to acquire a color image for texturing the object under surveillance or observation, therefore they are multimodal (bimodal) in nature (Liu et al., 2006; Liu and Stamos, 2007). There are a lot of works on 3D reconstruction of scenes (e.g., Stamos and Allen, 2000) and obstacle detection for robotics; we can also find quite some works on target detection and recognition using laser range finders (Mohottala, et al., 2009). Another type of laser sensor - a Laser Doppler Vibrometer (LDV) - is a long-range, non-contact acoustic measurement device to detect the speed of the target's vibration based on Doppler frequency shift. In fact, laser vibrometry has attracted attention for its use in many other applications, such as bridge and building inspection (e.g., Khan, et al., 2000), vehicle classification (Nedgård, 2005; Masagutov, et al., 2007), medical and screening applications (Lai, et al., 2008), and search and rescue scenarios. Therefore research in improving and utilizing this novel sensor will be beneficial to not only surveillance applications but also many other applications. The LDV sensors could either be point sensors (Zhu et al, 2007; Masagutov, et al, 2007; Lai et al, 2008) or array sensors (Nedgård, 2005).

2.1.4 Other Sensors and Modalities

Radar (Radio detection and ranging) can detect and range a target from a distance. For surveillance applications, radar range sensors are used in aid of other EO and/or IR sensors. The systems work by using radar range sensors to build up a ground truth map of the area to be monitored and set up range markers around a known central point. By

overlaying the range map with an EO and/or IR image and calibrating using rigorous on-site models, security zones are accurately set up within a meter. Sonar (sound navigation and ranging) is an acoustic type sensing technology that propagates sound ranging from low (infrasonic) to extremely high (ultrasonic) to detect objects, particularly vessels in underwater. Other sensors like ultrasound devices can also detect objects without physical contact. They work on principles that are similar to radar and sonar which evaluate attributes of a target by interpreting the echoes from radio or sound wave respectively. Another type of acoustic sensors frequently used in surveillance applications consists of microphone arrays. They are usually used to obtain the audio signals of the speakers. An interested work using microphones to track the speaker's position in videoconferencing and surveillance applications is presented by Zotkin et al. (2007). However, microphones are limited in a large distance and non-contact environment for surveillance applications.

2.2 Multimodal Surveillance System

In this section, we mainly emphasize two surveillance systems, video surveillance system and audio surveillance system. Computer vision techniques for visual surveillance tasks can be divided into three steps: 1) Low level processing, dealing with the extracting of salient simple features from a single image, such as edges, corners, homogenous regions, and curve fragments; 2) Intermediate level processing, dealing with the extraction of semantically relevant characteristics from one or more images, such as group features (structures), depth, and motion information; 3) High level processing, dealing with the

interpretation of the extracted information into object classes and activities. Most visual surveillance problems start with object detection and aim at segmenting regions corresponding to specific colors or shapes such as human skins or faces or whole bodies, or moving objects such as walking/running humans and moving vehicles from the static background. Interesting object features are then calculated to track the objects if either the sensor or the objects move. Vision-based surveillance algorithms have been extensively investigated (e.g., Foresti et al., 2005; Tian et al., 2008b). The underlying algorithms consist of methods ranging from simple background extraction algorithms to more complex methods such as optical flow methods (Shin et al., 2005).

Audio-based surveillance algorithms are always used to detect and recognize specific audio events. Those algorithms typically start with a supervised model or a training phase in which various features are extracted to obtain the signatures of different types of events. Then income sound signals are matched to the trained acoustic signatures to detect events (Clavel et al., 2005; Harma et al., 2005). However, supervised models learned for sound classes would only be able to detect suspicious activities known in advance. A hybrid solution is proposed by Radhakrishnan et al. (2005) that consists of two parts. The first part performs unsupervised audio analysis and that the second part performs analysis using an audio classification framework obtained from off-line training. Therefore it is capable of detecting new kinds of suspicious audio events (Radhakrishnan and Divakaran, 2006).

Video surveillance systems using only EO sensors offer a number of advantages over many other sensor modalities, particular in wide area detection. However, they have limitations: they can only obtain one type of characteristics, i.e., the appearance or the shape of an object. Thus, the addition of other types of sensor modalities can introduce an increase in the value of systems. Thermal/Infrared sensors may be one of the choices to acquire the temperature information of the target comparing to the surrounding. Similarly, audio only surveillance systems are good at detecting and recognizing suspicious events based on the acoustic signals, but they cannot track objects in the absence or the discontinuity of the sound. As a solution, visual sensors can be employed to detect and track objects, and then special events can be recognized using audio information. The integration of EO/IR and audiovisual data are further described in the next section.

2.3 Multimodal Data Fusion

Multimodal data fusion is the process of combing data from multiple sources in order to provide a better acceptable degree of robustness or more reasonable accuracy than using only one individual source. There are several advantages using multimodal data fusion. First, *different* information can be obtained from different sensor modalities. Second, reliability of the results can be improved by using *redundant* information. Third, errors rate can be reduced with *complementary* information. Data fusion can be performed at different processing levels either before or after the object classification (Figure 2.2). The fusion takes place before the classification stage is called *early integration*. The class

carries out fusion in the decision level is referred as *late integration*. There are also some methods fusing within the classification process, which can be called *intermediate integration*. Sometimes these integration levels also referred to in the literature as: premapping fusion, midst-mapping fusion, and postmapping fusion (Sanderson and Paliwal, 2004) (Figure 2.2).

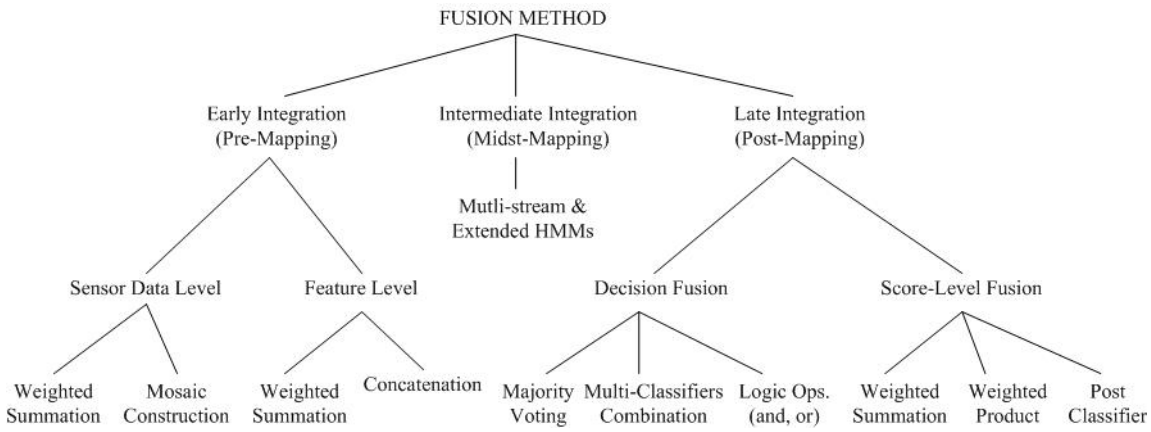


Figure 2.2 Multimodal Data Fusion (adapted from Sanderson and Paliwal, 2004)

2.3.1 Levels of Integration

There are three integration levels: early integration, intermediate integration and late integration. Early integration can happen at the sensor data level and/or the feature level. In the sensor data level (Hall and Llinas, 2001), the data from different sensors is combined. Two methods, weighted summation and mosaic construction, are commonly utilized to accomplish data fusion at sensor data level. In the *weighted summation* approach, the data is first normalized, and then combined to map to a common interval. When doing this, the assumption is that the data from multiple sources have been transformed into the same types of data and are aligned. *Mosaic construction* as a specific

vision technique is utilized to create one large image out of multiple images provided by several vision sensors. For doing this, these images need to be aligned either by a pre-calibration step or online image matching step. *Feature level fusion* is the combination of the features obtained from different sensor modalities. Joint feature vectors can be obtained by weighted summation after normalization, or by concatenating the feature vector from one modality to the feature vector from another different modality. However, the concatenation approach usually has high dimensionality, which can affect reliable training of a classification system (the "curse of dimensionality") (Theodoridis and Koutroumbas, 2008).

Intermediate integration is processed during the procedure of mapping the feature space into the decision space. It usually employs a single classifier that is responsible for both the data fusion and event classification. The architecture of the classifier should include the low-level intermodal elements and exploits the temporal dynamics contained in different stream so that the dimensionality problem and requirement of matching rates can be avoided. *HMM-based methods* are commonly used to compute the state probabilities of the stream components and to calculate the state-occupancy path which reflects the internal dynamics of the HMMs. In earlier work, single or two-stream HMMs (Potamianos and Graf, 1998) are carried out for the fusion and classification of bimodal speech via integrating the speaker's audio and lip motion information in a single state machine. Synchronized progression of both modalities has to be assumed. Currently, multi-stream and extended HMMs (Bengio, 2003; Potamianos et al., 2003; Aleksic et al., 2002, Chu & Huang, 2007) are commonly used for intermediate fusion in audiovisual

speech recognition. Multi-stream HMMs allows easy modeling of the reliability of the audio and visual stream and various levels of asynchronicity between them.

As stated earlier, late integration carries out fusion in the decision level. The final decision can be determined through the majority voting, logical operations, or combination of all classifiers. In *majority voting* (Radova and Psutka, 1997), the majority of the classifiers make the final decision. In *logical AND fusio*n, the final decision is made only if all classifiers reach the same decision whereas the decision can be made as soon as one of the classifiers is reached in logical OR fusion. A *multi-classifier combination method* combines all classifiers and makes a list of ranks. Not all decisions the classifiers have to be made, it is possible to have the level of score on each decision. Then all scores are combined utilizing weights using either weighted summation or weighted production. The weights are determined based on the discriminating ability of the classifier and the quality of the feature extraction. In a *postclassifier opinion fusion approach* (Sanderson and Paliwal, 2004), the likelihoods corresponding to each of N_C classes of interest, obtained utilizing each of the N_E available classifiers, are considered as features in the $N_C \times N_E$ dimensional space, where the classification of the resulting features is performed.

2.3.2 Multimodal Fusion Examples

Research in the fusion of visible and infrared imagery has received considerable attention in the past; however, fusion in *video* modalities (i.e., using continuous image sequences) for automatic surveillance is recent. In Torresane et al. (2004), the fusion of thermal infrared with visible spectral video, in the context of surveillance and security, is done by

building object correspondence. Davis and Sharma (2005) present a new contour-based background-subtraction technique using thermal and visible imagery for persistent object detection in urban settings. Recently Shah et al. (2010) have presented the fusion of infrared (IR) and visible surveillance images using the combination of wavelets and curvelets. There are also several pieces of work in thermal-visible video fusion for moving target tracking (Conaire et al.; 2006; Leykin et al., 2007; Krotoski and Trivedi, 2008; Zhao and Cheung, 2009).

Audio surveillance systems are not good at tracking moving objects due to the discontinuity of the sound. They are more or less used with the aid of video for the event detection and the object (human) recognition. Recent works on audio and video data fusion have been applied to speech processing (Hershey et al., 2004) and recognition (Chu and Huang, 2007). In surveillance applications, audio-visual integration has also been studied. In Cristani et al. (2006), the audio-visual foreground extraction for event characterization is presented. Both audio and visual information are analyzed by a standard background-foreground modeling. Online association and integration of audio and video information is performed. Thus the synchrony of foreground is assumed. In Vu et al. (2006), the audio-video surveillance for the automatic surveillance and public transportation is presented. Late integration strategy is then performed on audio and video events based on spatio-temporal reasoning. Similar work on late integration of audio and video on people fighting using decision (AND) based method is presented by Dedeoglu et al. (2008). Recently, Codec et al. (2010) proposed an autonomous vehicle classification and detection system based on audio-visual co-training using low-cost

consumer sensors that avoided the use of complicated calibration and expensive microphone arrays. Fusion is made during online learning by training two heterogeneous classifiers on a small amount of labeled data to co-train them on a continuous stream of unlabeled data to yield highly adaptive classifiers.

2.4 Motivations of our Approaches

For the multimodal sensing system design, we select a pair of PTZ camera and a LDV. The pair of PTZ camera can detect and track targets at a large distance, and further obtain the target distance. In general, human (or vehicle) detection mostly depends on visual information, whereas the audio modality is used as complementary information to discover and explain interesting activities in a scene. In some scenarios, however, audio conveys more significant information than video, for example, a human talking behind an object, or two people with similar appearance facing back against the camera. In the past, microphones or microphone arrays have been employed in audio-visual surveillance (Zotkin, et al, 2001; Maganti, et al, 2007; Gatica-Perez, et al, 2007; Codec et al, 2010), but they have the limitation of very short ranges. Furthermore, these types of sensors need to be fixed at pre-determined locations. If the targets move out of their sensing ranges, they will not be able to obtain any signals. A parabolic microphone can capture voice signals at a fairly large distance; however, when it points to the direction of the target, all the signals on the way are captured. A LDV is a non-contact acoustic sensor can detect voice signals at a large distance through the detection of vibration of a surface object near a sounding target. Therefore, they can be used to perform long-range multimodal

surveillance and monitoring by integrating visible and infrared video. In this thesis, we will show how those sensory components are integrated and calibrated for active control and adaptive sensing.

Data collected from multiple sensory components are always noisy and not aligned in a way that can be easily processed before performing feature level integration. They need to be associated and well represented to indicate the same objects. Then various features can be extracted for object classification. However, several environmental variations will significantly affect the accuracy of object classification. This will be even more the case for long-range object detection and inspection, where the sensors can only be set in a remote location. In vehicle detection and classification for applications such as traffic management and check-point vehicle inspection, the standpoints of and views from the sensors to a road could be constrained due to large distances for safety or installation reasons, and there could also be occlusions such as by tress and other facilities. We thus provide an effectively technique to solve these issues for a better classification performance. Furthermore, in feature level integration, large feature vectors may be created from multiple modalities. But not all features are equally important to make a good decision. The issue of what to fuse has been addressed at two different levels: feature modality selection and feature vector reduction. Many feature vector reduction techniques have been applied. Commonly used are principle components analysis (PCA), vector decomposition (SVD) and linear discriminant analysis (LDA). Since most previous work only focused feature vector reduction but few discussed feature modality selection, we will mainly focus on the feature modality selection. This also emphasizes the AIM-SP

framework that we proposed to learn representative feature modalities for adaptive multimodal sensors.

More importantly, we would like to provide a mechanism to tailor a multimodal sensor fusion system to a wide range of various tasks (human detection, vehicle detection, bridge monitoring, etc) using the same inference framework through optimal feature selection and ensemble classification learning. Many object detection problems can be formulated as classification problems. For example, for human detection in surveillance and search-rescue applications, the problem of human detection can be formulated as a two-class classification problem: human or no human. We would like to provide a systematic analysis of what features are selected, what scales are the best for a given tasks, and how heterogeneous, multimodal data are used in integrating those data, and how the selections of classifiers and features can be used for improving real-time sensing (smart data collection) and further for providing insights in new sensor designs.

Chapter 3

3 Multimodal Sensing and Adaptation

Remote object signature detection is becoming increasingly important in non-cooperative and hostile environments for many applications, such as wide-area surveillance, perimeter protection and search and rescue (Dedeoglu, et al, 2008; Li, et al, 2008). Although imaging and video technologies (including visible and IR) have had great advancement in object signature detection at a large distance, there are still many limitations in non-cooperative and hostile environments because of intentional camouflage and natural occlusions. Audio information, another important data source for target detection, can provide complementary information. For obtaining better performance of human tracking in a near to mediate range, Beal, and et al. (2003) and also Zou and Bhanu (2005) have reported the integrations of visual and acoustic sensors. By integration, each modality may compensate for the weaknesses of the other one. But in these systems, the acoustic sensors (microphones) need to be placed near the subjects in monitoring, therefore cannot be used for long-range surveillance. A parabolic microphone, which can capture voice signals at a fairly large distance in the direction pointed by the microphone, could be used for remote hearing and surveillance. But it is very sensitive to noise caused by the surroundings (i.e. wind) or the sensor motion, and all the signals on the way are captured. Therefore there is a great necessity to find a new type of acoustic sensor for long-range voice detection.

Laser Doppler Vibrometers (LDV) such as those manufactured by Polytec (2009) and Ometron (2009) can effectively detect vibration within two hundred meters with sensitivity in the order of 1μm/s. Larger distances could be achieved with the improvements of sensor technologies and the increase of the laser power while using a different wavelength (e.g. infrared instead of visible). In our previous work (Li, et al., 2006; Zhu, et al., 2007), we have presented very promising results in detecting and enhancing voice signals of people from large distances using a Polytec LDV. However, the user had to manually adjust the LDV sensor head in order to aim the laser beam at a surface that well reflects the laser beam, which was a tedious and difficult task. In addition, it was very hard for the user to see the laser spot at a distance above 20 meters, and so it was extremely difficult for the human operator to aim the laser beam of the LDV on a target in a distance larger than 100 meters. Of course human eyes cannot see infrared laser beams so it would be a serious problem if the LDV uses infrared. Also, it takes quite some time to focus the laser beam even if the laser beam is pointed to the surface. Therefore, reflection surface selection and automatic laser aiming and focusing are greatly needed in order to improve the performance and the efficiency of the LDV for long-range hearing.

Here, we present a novel multimodal sensing system, which integrates the LDV with a pair of pan-tilt-zoom (PTZ) cameras to aid the LDV in finding a reflective surface and focusing its laser beam automatically, and consequently the system captures both video and audio signals synchronously for target detection using multimodal information: in addition to video and audio, this sensing system can also obtain range information using the LDV-PTZ triangulation as well as stereo vision using the two cameras. The range information will

further add values to object signature detection in addition to the audio and video information, and improve the robustness and the detection rate of the sensor. The main contribution of this work is the collaborative operation of a dual-PTZ-camera system and a laser pointing system for long-range acoustic detection. To our knowledge, this is the first work that uses a PTZ stereo for automating the long-range laser-based voice detection. Meanwhile, the combination is a natural extension of the already widely used PTZ-camera-based video surveillance system towards multimodal surveillance with audio, video and range information. In order to acquire the audio and visual data synchronously and correctly, sensory components need to be calibrated and integrated properly. This step is important to ensure the reliability and usefulness of the obtained multimodal data, but this is not well studied before performing multimodal data integration. In this chapter, we will first present the components of the multimodal sensor platform, and then describe the calibration procedure. Finally we will present the active sensing and adaptive control using the multimodal sensor systems.

The rest of this chapter is organized as follows: Section **Error! Reference source not found.** presents some background and related work. Section 3.2 describes an overview of our vision-aided automated vibrometry system. Section 3.3 discusses the calibration issues among the multimodal sensory components. Section 3.4 shows the algorithms for feature matching and distance measuring using the system. Section 3.6 describes the adaptive and collaborative sensing approach. Experimental results and conclusions are provided in Section 3.7 and Section **Error! Reference source not found.**, respectively.

3.1 LDV for Remote Acoustic Sensing

3.1.1 Principle of LDV-Based Hearing

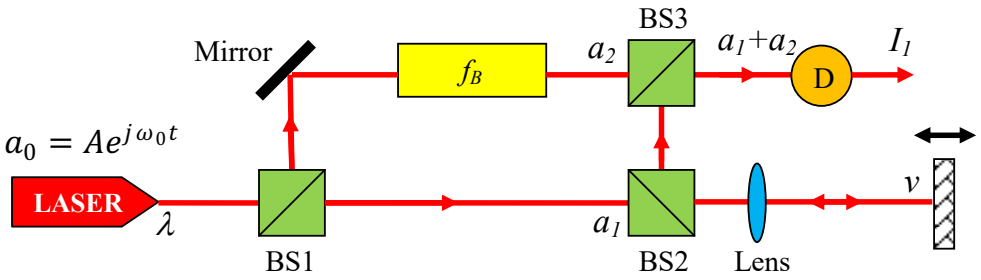


Figure 3.1 Principle of the Laser Doppler Vibrometer (LDV)

The laser Doppler vibrometer (LDV) works according to the principle of laser interferometry. Measurement is made at the point where the laser beam strikes the structure under vibration. In the Heterodyning interferometer (Figure 3.1), a coherent laser beam is divided into object and reference beams by a beam splitter BS1. The object beam strikes a point on the moving (vibrating) object and light reflected from that point travels back to beam splitter BS2 and mixes (interferes) with the reference beam at beam splitter BS3. If the object is moving (vibrating), this mixing process produces an intensity fluctuation in the light as

$$I_1 = \frac{1}{2}A^2\left\{1-\cos\left[2\pi\left(f_B+\frac{2v}{\lambda}\right)t\right]\right\} \tag{3.1}$$

where I_1 is light intensity; A is the amplitude of the emitted wave; f_B is modulation frequency of the reference beams; λ is the wavelength of the emitted wave; v is the object's velocity and t is observation time. A detector converts this signal to a voltage

fluctuation. And from the fluctuating of light patterns, the velocity of object can be decoded by a digital quadrate demodulation method (Scruby & Drain, 1990). An interesting finding of our study is that most objects vibrate while wave energy (including that of voice waves) is applied on them. Although the vibration caused by the voice energy is very small compared with other vibration, it can be detected by the LDV, and be extracted with advanced signal filtering. The relation of voice frequency *f*, velocity *v* and magnitude *m* of the vibration is

$$v = 2\pi f m \tag{3.2}$$

As seen from the above principle of the LDV, There are three requirements to be considered in order to use the LDV to measure the vibration of a target caused by sounds:

(1) An appropriate surface close to the sounding target with detectable vibration and good reflection index;

(2) The focus of the LDV laser beam on the refection surface, otherwise very weak reflection signals are obtained due to the scattering of coherent light and path length differences;

(3) A necessary signal enhancement process to filter out the background noise and the inherent noise of the LDV.

In close-range and lab environments, it is not a serious problem for a human operator to find an appropriate reflective surface, focus the laser beam and acquire the vibration signals. But at a large distance (from 20 meters to hundred meters), the manual process

becomes extremely difficult because it is very hard for a human operator to aim the laser beam to a good reflective surface. Also, it takes quite some time to focus the laser beam even if the laser beam is pointed to the surface. Therefore, there are great unmet needs in facilitating the process of surface detection, laser aiming, laser focusing, and signal acquisition of the emerging LDV sensor, preferably through system automation.

3.1.2 Related Work on Acoustic Sensing

Acoustic sensing and event detection can be used for audio-based surveillance, including intrusion detection (Zieger, et al., 2009), abnormal situations detection in public areas such as banks, subways, airports, and elevators (Clavel, et al., 2005; Radhakrishnan, et al., 2005). It can also be used as a complementary source of information for video surveillance and tracking (Cristani, et al., 2007; Dedeoglu, et al., 2008). In additional to microphones, a Laser Doppler Vibrometer (LDV), as another type of acoustic sensors, is a novel type of measurement device to detect a target's vibration in a non-contact way, in applications such as bridge inspection (Khan, et al., 1999), biometrics (Lai, et al., 2008), and underwater communication (Blackmon & Antonelli, 2006). It has also been used to obtain the acoustic signals of a target (e.g., a human or a vehicle) in a large distance by detecting the vibration of a reflecting surface caused by the sound of the target next to it (Zhu, et al., 2005; Li, et al., 2006; Zhu, et al., 2007; Wang, et al, 2011a). The LDVs have been used in the inspection industry and other important applications concerning environment, safety and preparedness that meet basic human needs. In bridge and building inspection, the non-contact vibration measurements for monitoring structural defects eliminate the

need to install sensors as a part of the infrastructure (e.g., Khan, et al., 1999). In security and perimeter applications, an LDV can be used for voice detection without having the intruders in the line of the sight (Zhu, et al., 2005). In medical applications, an LDV can be used for non-contact pulse and respiration measurements (Lai, et al., 2008). In search and rescue scenarios where reaching humans can be very dangerous, an LDV can be applied to detect survivors which are even out of visual sight. Blackmon and Antonelli (2006) have tested and shown a sensing system to detect and receive underwater communication signals by probing the water surface from the air, using an LDV and a surface normal tracking device.

However, in most of the current applications, such systems are manually operated. In close-range and lab environments this is not a very serious problem. But in field applications, such as bridge/building inspection, area protection or search and rescue applications, the manual process takes a very long time to find an appropriate reflective surface, focus the laser beam and get a vibration signal; more so if the surface is at a distance of 100 meters or more. A vision-aided LDV system can improve the performance and the efficiency of the LDV for automatic remote hearing. In this work, we improved the flexibility and usability of the vision-aided automated vibrometry system from our previous design with a single PTZ camera (Qu, et al., 2010) to the current design with a pair of PTZ cameras and by providing adaptive and collaborative sensing.

3.2 Vision-Aided Automated Vibrometry: System Overview

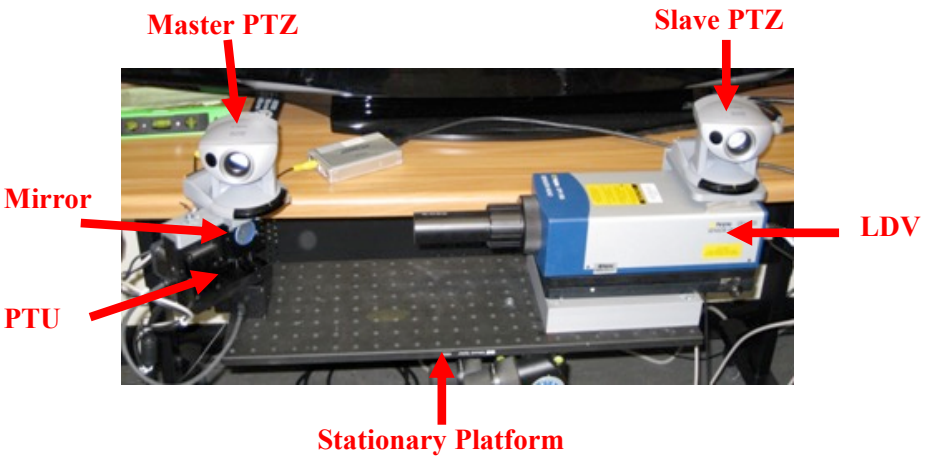


Figure 3.2 The multimodal sensory platform

The system consists of a single point LDV sensor system, a mirror mounted on a pan-tilt unit (PTU), and a pair of pan-tilt-zoom (PTZ) cameras, one of which is mounted on the top of the PTU (Figure 3.2). The sensor head of the LDV uses a helium-neon laser with a wavelength of 632.8 nm and is equipped with a super long-range lens. It converts velocity of the target into interferometry signals and magnitude signals, and send them to the controller of the LDV that are controlled by the computer via an RS-232 port. The controller processes signals received from the sensor head of the LDV, and then output either voltage or magnitude signals to the computer using an S/P-DIF output. The Polytec LDV sensor OFV-505 and the controller OFV-5000 that we use in our experiments can be configured to detect vibrations under several different velocity ranges: 1 mm/s/V, 2 mm/s/V, 10 mm/s/V, and 50 mm/s/V, where V stands for velocity. For voice vibration of a basic frequency range from 300 to 3000 Hz, we usually use the 1mm/s/V velocity range. The best resolution is 0.02 μm/s under the range of 1mm/s/V according to the manufacture's specification with retro-reflective tape treatment. Without the retro-

reflective treatment, the LDV still has sensitivity on the order of 1.0 μm/s. This indicates that the LDV can detect vibration (due to voice waves) at a magnitude in nanometers without retro-reflective treatment; this can even get down to picometer with retro-reflective treatment.

The LDV sensor head weights about 3.4 kg; this is the major reason that a mirror mounted on the PTU is used in our system to reflect the laser beam to freely and quickly point it to various directions in a large field of view. The laser beam point to the mirror at the center of the panning tilting of the PTU. The vision component consists of a pair of Canon VC-C50i (26x) PTZ cameras with one mounted on the top of the PTU, which is called the *master PTZ* since it is the main camera to track the laser beam, and another one mounted on the top of the LDV, which is called the *slave PTZ*. Each PTZ camera (Canon VC-C50i) has a 720×480 focal plane array and an auto-iris zoom lens that can change from 3.5mm to 91mm (26× optical power zoom). The pan angle of the PTZ is ±100° with rotation speed 1° to 90° per second and the tilt angle of it is from -30º to +90º with rotation speed 1° to 70° per second. The PTU is the model PTU-D46-70 of Directed Perception, Inc. It has a pan range from -159° to +159º and a tilt range from -47° to +31°. Its rotation resolution is 0.013° and max rotation speed is 300°/s. The reason to use zoom cameras is to detect targets and to assist the laser pointing and focusing at various distances. However, at a long distance, the laser spot is usually hard to be seen by the cameras, either zoomed or with wide views, if the laser is unfocused or not pointed on the right surface. Therefore, the master PTZ camera is used to rotate synchronously with the reflected laser beam from the mirror in order to track the laser spot. Although the laser point may not be observed

from the master PTZ, we always control the pan and tilt angles of the master PTZ camera so that its optical axis is in parallel to the reflected laser beam, and therefore the laser spot is always close to the center of the image. Then, the master PTZ camera and the slave PTZ form a stereo vision system to obtain the distance to focus the laser spot as well as guide the laser to the right surface for acoustic signal collection. The baseline between of the two PTZ cameras is about 0.6 meters for enabling long-range distance measurements. In order to obtain the distance from the target surface to the LDV, the calibration among the two PTZ cameras and the LDV is the first important step, which will be elaborated in next section before the discussion of our method for distance measurement.

3.3 System Calibration: Finding Parameters among the Sensor Components

There are two stereo vision components in our system: stereo vision between the two PTZ cameras, and stereo triangulation between the slave PTZ camera and the mirrored LDV laser projection. The first component is used to obtain the range of a point in a reflective surface by matching its image projection (x, y) in the master camera to the corresponding image point (x', y') in the salve camera. The second component is mainly used to obtain the pan (α) and tilt (β) rotation angles of the PTU so that the LDV points to the image point (x, y) in the master image. Before determining the distance, several coordinate systems corresponding to the multi-sensory platform (in Figure 3.2) is illustrated in Figure 3.3 (left): the master PTZ camera coordinate system (S_c), the slave PTZ camera coordinate system ($S_{c'}$), the LDV coordinate system (S_L), and the PTU coordinate system (S_u).

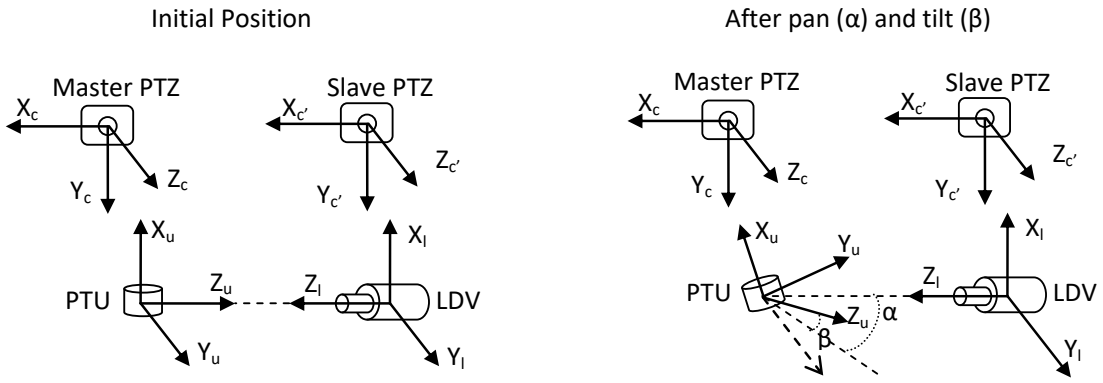


Figure 3.3 Coordinate systems of the multimodal platform

We assume the mirror coordinate system is the same as the PTU coordinate system since the laser will point to the mirror at the origin of the PTU system. The mirror normal direction is along the Z_u axis and initially points to the outgoing laser beam along Z_l. In order to always actually track the reflected laser beam visible or invisible (by having the optical axis of the master PTZ parallel to the reflected laser beam), the master PTZ not only rotates the same base angles with the PTU, α and β, which are the pan and tilt angles of the PTU around the X_u and Z_u axes, but also undergo additional pan and tilt rotations (α' and β') around the Y_c and X_c axes. We will explain in details how to determine these angles later.

The stereo matching is performed after the full calibration of the stereo component of the two PTZ cameras, and that between the slave PTZ camera and the "mirrored" LDV. Given a selected point on a reflective surface in the image of the master camera, we first find its corresponding point in the image of the slave camera, meanwhile calculating the pan and tilt angles of the PTU and the master and slave PTZ camera so that the laser spot is right under the center of the image of the master PTZ camera; the offset to the center

is a function of the distance of the surface to the sensor system. The farther the surface is, the closer is the laser spot to the center. The distance from the target point to the optical center of the LDV is estimated via the stereo PTZ and then used to focus the laser beam to the surface.

3.3.1 Calibration of the two PTZ cameras

The calibration between the two PTZ cameras is carried out by estimating both the intrinsic and extrinsic parameters of each camera on every possible zoom factor when the camera is in focus, using the same world reference system. We use the calibration toolbox by Bouguet (2008) to find a camera's parameters under different zoom factors. We have found that the estimated extrinsic parameters do not change much with the changes of zooms. However, the focal lengths of the cameras increase nonlinearly with the changes of different zooms therefore we have calibrated the camera under every possible zoom. Also note that the focal lengths of two PTZ cameras may not be same under the same zoom factor. In order to achieve similar fields of view (FOVs) and to ease the stereo matching between two images, the correct zoom of the slave PTZ camera corresponding to the actual focal length of the master PTZ camera should be selected. After the calibration, we obtain the effective focal lengths and image centers of the two cameras under every zoom factor k, and the transformation between the two cameras, represented by R and T:

$$P_{C'} = RP_c + T \tag{3.3}$$

where P_c and $P_{c'}$ are the representations of a 3D point in the master and slave PTZ coordinate systems (S_c and $S_{c'}$), respectively.

3.3.2 Calibration of the slave camera and the LDV

Since the intrinsic parameters of the slave PTZ camera have been obtained previously, we only need to estimate the extrinsic parameters characterizing the relation between the LDV coordinate system (S_L) and the slave PTZ camera coordinate system ($S_{C'}$), defined as:

$$P_L = R_{C'} P_{C'} + T_{C'} \tag{3.4}$$

where P_L and $P_{C'}$ represent the coordinates of a 3D point in S_L and $S_{C'}$, respectively. The $R_{C'}$ and $T_{C'}$ are the rotation matrix and translation vector between S_L and $S_{C'}$. The relation between the points in the LDV and the PTU systems is defined as:

$$P_L = R_U P_U + T_U \tag{3.5}$$

where R_U and T_U are the rotation matrix and translation vector between S_L and the PTU coordinate system. According to the principle of mirroring, the relation between the mirrored LDV coordinate system (S_{ML}, not shown in Fig. 3.3) and the PTU is defined as:

$$P_{ML} = R_U R_{LR} P_U + T_U \tag{3.6}$$

where P_{ML} are the 3D point representations in the S_{ML}. The R_{LR} is the rotation matrix that converts a right hand coordinate system to a left hand coordinate system, defined as:

$$R_{LR} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \tag{3.7}$$

From Eq. 3.5 and 3.6, the relation between the LDV coordinate system (S_L) and the mirrored LDV coordinate system (S_{ML}, not shown in Fig. 3) is defined as:

$$P_L = R_U R_{LR} R_U^T (P_{ML} - T_U) + T_U \tag{3.8}$$

where P_L and P_{ML} are the 3D point representations in the S_L and the S_{ML}, respectively, and R_U and T_U are the rotation matrix and translation vector between S_L and the PTU coordinate system.

Then the extrinsic parameters are estimated by combining Eq. (3.4) and Eq. (3.8), as

$$R_{C'} P_{C'} = R_U R_{LR} R_U^T (P_{ML} - T_U) + (T_U - T_{C'}) \tag{3.9}$$

For the calibration between the LDV and the slave PTZ, the LDV laser beam is projected at pre-selected points in a checkerboard placed at various locations/orientations. Because both the variables P_{ML} - T_U and T_U -$T_{C'}$ are not independent in Eq. (3.9), the distance between the fore lens of the LDV and the laser point on the mirror is estimated initially. Also, to avoid the complexity of the nonlinear equation we assume the initial rotation matrix is identify matrix which can be manually adjusted by pointing both cameras parallel to the same direction. Then this initial distance and initial rotation matrix can be refined iteratively. Giving n 3D points, 3n linear equations that include n+14 unknowns are constructed using Eq. (3.9). Therefore, at least 7 points are needed. More details can be found in Appendix A.

3.4 Stereo Vision: Feature Matching and Distance Measuring

3.4.1 Stereo Matching

After calibration, distance of a point can be estimated when the corresponding point in the slave image of a selected point in the master image is obtained. In the master camera, a target point can be selected either manually or automatically. We assume that both left and right images can be rectified given the intrinsic matrices for both cameras and the rotation matrix and translation vector. So, given any point $(x,y,1)^T$ in original (right) image, the new pixel location $(x',y',1)^T$ in rectified right image is $R'_r\ (x,y,1)^T$. To simplify the task radial distortion parameters are ignored. The rectified matrices for both cameras (virtually) make both camera images plane the same plane. Thus, the stereo matching problem turns into a simple horizontal searching problem since all epipolar lines are parallel. For example, in Figure 3.4, the right image is captured by the master PTZ camera and the left image by the slave PTZ camera. The same target points are shown in white circles. The numbers above the white circles show the pan and tilt angles of the PTU in order to point the laser beam to the target point. The epipolar line is shown in green line cross both images. Note that due to the calibration error, the corresponding point may not be exactly on the epipolar line. To solve the problem, a small search window is used to match the region around the selected point with a small range in the vertical direction as well. Since we are only interested in the selected point on a particular reflective surface, this step is very fast.

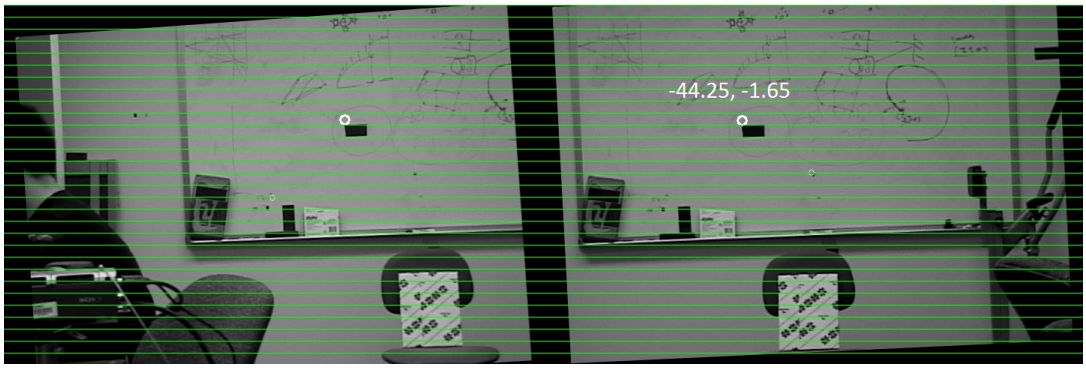


Figure 3.4 Stereo matching of the corresponding target point.

3.4.2 Distance Measuring

Once two corresponding points lying on the same horizontal epipolar line are identified, the distance can be calculated based on the triangulation using the baseline (B) of two rectified cameras. The relation between B and the range (*D*) of the target surface represented in the master camera system is defined as:

$$\frac{B}{D} = [x_r \ -x_l]\left[\frac{1}{F_{xr}} \ \frac{1}{F_{xl}}\right]^T \tag{3.10}$$

where x_r and x_l are the x coordinates of the selected point in the right and left image, F_{xr} and F_{xl} are the focal lengths of the two PTZ cameras. Ideally, both PTZ cameras should have the same focal length after adjusting their zoom factors.

The calibration result of the slave camera and the LDV is mainly used to determine the pan (α) and tilt (β) angels of the PTU in order to direct the laser beam to the selected point. The conventional triangulation method (Trucco & Verri, 1998) is used to match the ray from the optical center of the PTZ to the ray of the reflected laser beam. Then the

corresponding pan and tilt angels are estimated. Fig. 3.4 shows an example of the calculated pan and tilt angels (on the right image) corresponding to the point (in white circle) in the left image. Then, giving the pan and tilt rotations of the PTU and knowing the corresponding 3D point in the slave camera system as

$$P_{C'} = R'[P_{C'X}, P_{C'Y}, D]^T \tag{3.11}$$

where R' is the pan and tilt rotation of the slave PTZ. Initially it equals identity matrix if the slave PTZ camera is in its initial pose when it was calibrated. The estimated LDV distance $D_L = ||P_{ML}||$ can be then defined based on Eq. (3.9) that will be used for focusing the laser beam to the target point.

3.5 LDV Focus Step and Distance Relation

The calibration between the LDV and one of the PTZ cameras allows us to obtain the distance from the target to the lens of the LDV so that we can find the focus steps for the LDV quickly. The method for automatic fast focusing based on the distance measurement can be found in (Qu et al., 2011). Here for completion, we will briefly layout the relation between the distance and LDV focus step in order to achieve automatic fast focusing.

According to Gaussian lens equation, the relationship between the distance from the target (i.e. the reflective surface) to the lens D, and the distance from the lens to the image d is defined as:

$$\frac{1}{D} + \frac{1}{d} = \frac{1}{f_L} \tag{3.12}$$

Where f_L is the effective focal length of the lens of the LDV. A super-long range lens OFV-SLR (f_L = 200 mm) is used in the LDV, and the possible stand-off distance *D* of the target is from 1.8 meters to over 300 meters. While the focal length f_L is constant when the target distance *D* changes, the image distance of the LDV has to change for obtaining a focused image of the laser point. In the LDV, this is achieved by changing the focus step S (from 0 to 3300 digital steps).The relation of the image distance and the target distance can be calculated by Eq. 3.12. Due to the lack of the intrinsic parameters of the LDV, particularly the relation between the image distance and the focus steps (0 – 3300), we calibrate the relation experimentally. We measure the distances between fore lens of the LDV and reflective surfaces (targets) at various distances (from 2.13 to 200 meters), meanwhile acquiring the focus step values (from 893 to 2962) when using the built-in automatic focus function of the LDV to achieve laser beam focusing. Those corresponding values between the distances and the focus steps for the Polytec OFV-505 LDV are shown as red circles in Figure 3.5. The fitted curve is shown in black line on top of the measured data in red circles. Given the data we have measured, the fitted curve applies from 1.54 meters to 300 meters for the distance, and from 0 to about 3000 for the steps. This is consistent with the manufacture's specification

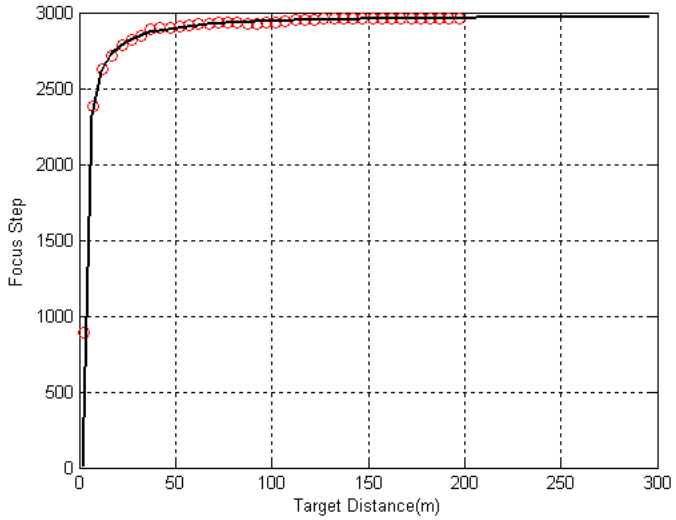


Figure 3.5 Focus-step and distance relation (The fitted curve is shown in black line on top of the measured data in red circles)

Theoretical, given the distance of the reflective surface, the corresponding LDV focus step (within the range from 0 to 3300 for the LD we used) can be calculated. However, there is problem of the step-distance measurement is not fine enough for accurate focusing. Therefore, an automatic multi-scale focusing algorithm based on the measured distance and LDV signal return level is further applied; Details of the algorithm can be found in our previous paper (Qu, et al., 2010).

3.6 Adaptive and Collaborative Sensing

The overall goal of this system is to acquire meaningful audio signatures with the assistance of video cameras by pointing and focusing the laser beam to a good surface. However, a target location either manually or automatically selected may not return signals with a sufficient signal-to-noise ratio (SNR). Then a reselection of new target points

is required. Figure 3.6 shows the basic idea of adaptive sensing of adaptively adjust the laser beam based on the feedback of its returned signal levels.

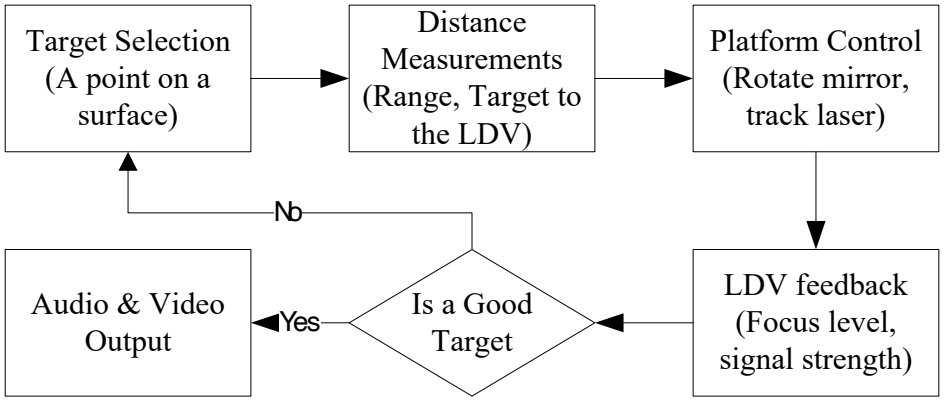


Figure 3.6 Flow chart of adaptive sensing for laser pointing and tracking for audio and video signature acquisition

The stereo matching here is used for obtained the target distance to the system platform, and then we can automatically focus the laser point to the selected target. This involves the following procedures.

First, a point on a surface close to a designated target is selected either manually or automatically.

Second, the target range and the distance from the point to the optical center of the LDV are measured.

Third, the laser spot is moved to the new location, and the master PTZ camera is rotated synchronously to put the laser spot in the center of images.

Fourth, the laser beam of the LDV is automatically and rapidly focused based on estimated distance and the signal levels, as we did in (Qu, et al., 2010).

If the selected target point does not have sufficient good returning signals for voice detection, we need to reselect new target points. If the target point is good enough, we can use it to record the audio signature as well as video signatures. In this procedure, there are two key issues need to be emphasized. First, what is a good surface and how to select a surface? Second, how to align the laser beam with the optical center of the camera accurately?

3.6.1 Surface Selection

The selection of reflection surfaces for LDV signals is important since it is a major factor that determines the quality of acquired vibration signals. There are two basic requirements for a good surface: *vibration to the voice energy and reflectivity to the helium-neon laser*. We have found that almost all natural objects vibrate more or less with normal sound waves. Therefore, the key technique in finding a good reflection surface is to measure its reflectivity. Based on the principle of the LDV sensor, the relatively poor performance of the LDV on a rough surface at a large distance is mainly due to the fact that only a small fraction of the scattered light (approximately one speckle) can be used because of the coherence consideration. A stationary, highly reflective surface usually reflects the laser beam of the LDV very well. Unfortunately, the body of a human subject does not have such good reflectivity to obtain LDV signals unless (1) it is treated with retro-reflective materials; and (2) it can keep still relative to the LDV. Also, it is hard to

have a robust signal acquisition on a moving object. Therefore, background objects nearby to the interested target are selected and compared in order to detect useful acoustic signals. Typically, a large and smooth background region that has a color most close to red is selected for the LDV pointing location.

3.6.2 Laser-Camera Alignment

Figure 3.7 Two examples of laser point tracking.

The next issue is to how to automatically aim and track the laser spot, especially for long range detection. The laser spot may not be observable at a long range particularly if it is not focused or it does not point on the surface accurately. We solve this problem by keeping the reflected laser beam always in parallel to the optical axis of the maser PTZ camera so that the laser spot is right under and very close to the center of the master image. Figure 3.7 shows a typical example of a laser spot (in red spot) that is right below the image center (in yellow circle) in few pixels. Both images in Figure 3.7 show the same cropped size (240x160) around the image center (yellow circle) with focused laser spot close to it (red-white dot). The laser point on a white board (indoor) is about 6 meters in the left image. We make the ray from the optical center of the master PTZ camera parallel to the reflected laser beam by rotating the PTZ camera with the PTU synchronously (since the PTZ is mounted on the PTU), then with additional PTZ camera rotations.

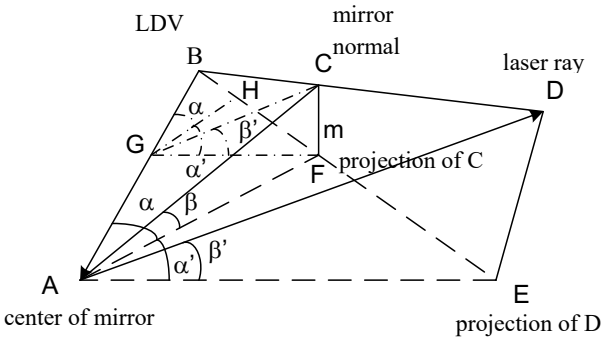


Figure 3.8 Geometric model of laser beam from the LDV (BA) and its reflected laser ray (AD) after the pan (α) and tilt (β).

The main issue now is how to obtain the additional pan (α') and tilt (β') angles of the PTZ camera given the pan (α) and tilt (β) angles of the PTU. Figure 3.8 shows the relationship between the outgoing laser beam from the LDV ($\overrightarrow{BA}$) and the reflected laser ray ($\overrightarrow{AD}$), with the mirror normal ($\overrightarrow{AC}$). By projecting the reflected ray and the mirror normal on the YZ plane (in both the PTU and the LDV coordinate systems in Figure 3.3), as $\overrightarrow{AE}$ and $\overrightarrow{AF}$ respectively, we see that the angle $\angle BAF$ is α and the angle $\angle FAC$ is β. Now the camera optical axis is parallel to the mirror normal AC. Two additional angles are defined in the figure, the pan angle α' as the angle $\angle FAE$, and the tilt angle β' as the angle $\angle EAD$. If the master PTZ camera (mounted on top of the PTU) is tilted back by - β, then its optical axis will be parallel to AF. Therefore, by further panning the PTZ by the angle α' and tilting the PTZ by the angle β', its optical axis will be in parallel with the reflected laser ray AD.

Define a helping line GC parallel to AD, we can easily solve the additional pan (α') and tilt (β') based on triangulation. The detailed derivation can be found at Appendix B. As a result, the pan angle (α') is

$$\alpha' = \tan^{-1}\left(\frac{\tan\alpha}{\cos 2\beta}\right) \tag{3.13}$$

and the tilt angle (β')

$$\beta' = \sin^{-1}(\sin 2\beta * \cos\alpha) \tag{3.14}$$

3.7 Experimental Results

In this section, we provide some results on distance measuring, surface selection, auto-aiming using laser-camera alignment, and surface focusing and listening using our multimodal sensory system.

3.7.1 Distance Measuring Validation

This experiment is used to verify the accuracy of the calibration among sensory components. Therefore, the test is performed under controlled environments, inside a lab room (with distances up to 10 meters) and in the corridor of a building (with distances up to 35 meters). Note that the camera's focal lengths do not increase linearly with the change of the zoom levels. In order to perform accurate distance measurement on a large distance, we calibrated the focal lengths under different zoom factors. Figure 3.9 shows the focal lengths (in both x-, y- directions) of the main PTZ camera and the slave PTZ camera.

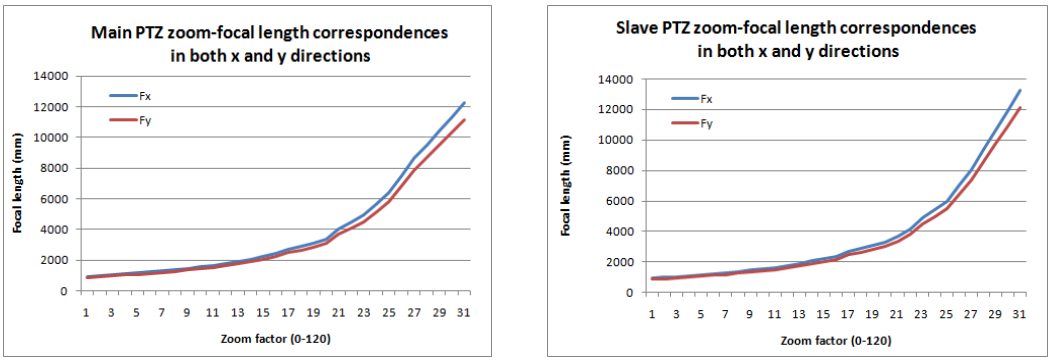


Figure 3.9 Calibrated focal lengths of the master PTZ camera and the slave PTZ camera under different zooms.

Next we verified the correctness of calibration parameters, especially with changes of the focal lengths of each camera. We used the same feature point on a check board pattern at various distances. At each zoom level, the distance from the check board to the platform was manually obtained as the ground truth, and then we used the calibrated parameters to estimate the distance at that zoom level. Figure 3.10 shows the comparison of the true and estimated distances under various zoom factors, which has an average relative error of 6%. The accuracy is sufficient for performing the adaptive focus of the LDV sensor.

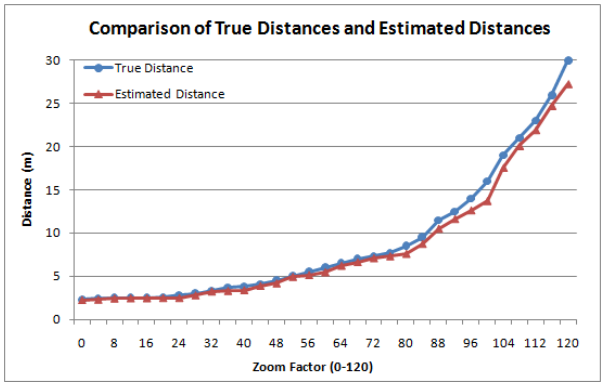


Figure 3.10 The comparison of true distances and estimated distances under various zoom factors

3.7.2 Surface Selection

In this experiment (Figure 3.11), several interested target points are automatically selected in the segmented regions close the human target in an image of the master PTZ. Figure 3.11 shows the image captured at a corridor, the cropped (320x240) original image (under zoom factor 48) with a target (in red rectangle) is shown on left. On right, interested target points close to the human target in the segmented regions are selected and labeled (as L1-L9). The distance of the camera is about 31 meters. Note that a static target object such as human or vehicle can be easily detected using histograms of oriented gradients (HOG) (Dalal & Triggs, 2005) in the image. If a target is moving, then frame difference can be used to separate the target from the background surfaces. Then conventional color segmentation can be performed. The region centroid points close to the center of the target can be selected to point the laser.

Figure 3.11 Surface selection in a segmented image of a 31 meters corridor.

3.7.3 Auto-Aiming using Laser-Camera Alignment

When an interested surface point is select, the master camera is centered to that point; then the laser-camera alignment technique automatically aim the laser spot close to or

right below the image center in focus using the calculated distance via stereo vision of the two PTZ cameras. Here we test our system in two environments, one is indoor (Figure 3.12) and another is outdoor (Figure 3.13).

Figure 3.12 Indoor auto aiming using laser-camera alignment.

The indoor experiment is performed at the corridor about 30 meters. A metal box on a chair is placed on a fixed location at about 9 meters. We manually selected three surfaces points, the points on the metal box, metal door handler and extinguisher metal box. The laser spots can be clearly observed in the images that are close to the image centers with pixel errors of 2.3, 5.2, and 5.5 (from left to right in Figure 3.12). In Figure 3.12, the yellow circles in the cropped images show the image center of the original image. The calculated distances are 8.9, 10.6, and 26.6 meters with corresponding true distances at 9.0, 11.0, and 28 meters

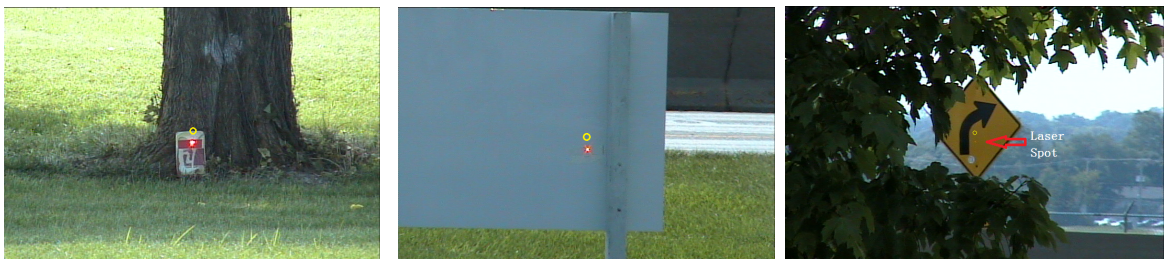


Figure 3.13 Outdoor auto aiming using laser-camera alignment

The outdoor experiment is performed near a highway when the sensor platform has a standoff distance of about 60 meters from the highway. In Figure 3.13, three sample surface targets to the side of the highway close to the sensor platform are selected, a metal box under a tree (45.4 meters), a poster with a tape (45.5 meters), and a right turn sign (53.3 meters). All images are zoomed so that both the image centers (in yellow circles) and the laser spots (in red) right below are visible. The average pixel difference between the laser spot to the image center for the three examples is 6 pixels.

3.7.4 Surface Focusing and Listening

The experimental results related to the distance measuring, surface selection and laser pointing for those labeled points (in Figure 3.11) are presented in Table 3.1. The estimated camera distance (D) are listed in column 3 with the "ground truth" data (D*) at column 6. The LDV distance (D_L) in column 4, the distance from the target point to the optical center of the LDV, is calculated based on the pan and tilt angles of the PTU. Base on that, the focus step (in the range of 0 to 3300) in column 5 is determined and the laser beam is focused in about 1 second for each point. For comparison, the focus step using the full range searching takes 15 second, and is presented in column 8. The signal returning levels (0 to 512) in column 6 can be used to determine what the best point is among the candidates for audio acquisition. As a result, the metal box (L7) has the strongest signal return level so that it is selected for the voice detection. Note that all selected surfaces do not have retro-reflective tape treatment.

Table 3.1 Surface selection, laser pointing and focusing

L#	Surface	Measurements				Ground Truth		
		D (m)	D_L(m)	Step	Level	D* (m)	Step	Level
L1	Floor	26.56	27.14	2642	10	27.74	2581	11
L2	Chalkboard	28.30	28.88	2750	31	27.74	2764	22
L3	Wall	27.67	28.25	2732	12	30.63	2734	12
L4	Wall	28.62	29.20	2734	11	30.63	2734	12
L5	Wall	29.67	30.26	2786	12	30.63	2845	14
L6	Mirror	27.90	28.56	2758	12	30.63	2757	12
L7	Metal box	29.67	30.26	2745	118	30.32	2839	121
L8	Side wall	21.10	21.60	2391	9	23.16	2410	10
L9	Wall	30.85	31.43	2410	11	30.63	2757	11

The experiment results of the focus positions and signal levels of the outdoor surface targets (in Figure 3.13) are shown in Table 3.2. According to the signal return levels at the last column, the surface of the metal box under a tree should be selected as the best listening target. In addition, the poster with tape is also a good listening surface with moderate signal level. Therefore it can be used as a substitute for the first one with some signal enhancing treatments, such as amplifying, noise removal, and filtering. Unfortunately the right turn sign does not provide sufficient signal returns.

Table 3.2 Focus positions and signal levels of three outdoor surfaces

No	Target	Distance	Focus Position	Signal Level
001	Box under a tree	45.4m	2890	285
002	Poster with tape	45.5m	2890	116
003	Right turn Sign	53.3m	2904	14

3.8 Concluding Remarks

In this chapter, we present a dual-PTZ camera based stereo vision system for improving the automation and time efficiency of LDV long-range remote hearing. The close-loop adaptive sensing using the multimodal platform allows us to determine good surface points and to quickly focus the laser beam based on target detection, surface point selection, distance measurements, and LDV signal return feedback. The integrated system greatly increases the performance of the LDV remote hearing and therefore its feasibility for audio-visual surveillance and long-range other inspection and detection applications. Experimental results show the capability and feasibility of our sensing system for long range audio-video-range data acquisition.

Chapter 4

4 Multimodal Data Representation and Processing

The calibrated system allows us to acquire both audio and video data synchronously, as well as the range information and the system configurations (such as zoom, pan and tilt parameters). However, it is still a challenging task to automatically extract, label and integrate multimodal data for the recognition and classification moving targets (e.g., humans, vehicles). In this chapter we will mainly focus on moving vehicles in uncontrolled traffic scenes for vehicle classification, check-point inspection and traffic analysis. The same principle could be used for other moving targets. In this work, we first represent both visual and audio data in a multimodal temporal panorama (MTP) (Wang, et al., 2011b), which shows detection, motion, and acoustic information simultaneously. The MTP provides a very effective user interface to visualize and analyze the alignment of the video and acoustic information of passing-by vehicles, thus facilitating the joint detection and classification of vehicles using both visual and audio information. It provides:

1) multi-modal information including visual presentation from a panoramic view image, motion presentation from an epipolar plane image, and acoustic information from an audio wave scroll;

2) real time detection, reconstruction of the vehicles' visual appearances, synchronized with their acoustic signatures; and

(3) a very effective user interface for training data labeling in both video and audio domains.

In addition, a robust vehicle reconstruction algorithm is developed using both panoramic view images and epipolar plane images (Wang & Zhu, 2012a). The reconstructions are useful since the vehicles may be occluded by other stationary objects, such as bushes, trees, parked vehicles or others. Motion blur can also be removed after reconstruction. In addition, all vehicles have the same side views that can improve the recognition and classification performance while keeping the classifier simple.

There are a number of advantages of this work. Since the generation of the MTP is done in real time, the reconstruction takes place immediately after a vehicle is detected. Second, audio information is used to remove some false detecting targets before reconstruction. Third, a multimodal dataset of different types of vehicles are generated automatically. Last, the classification of the reconstructed vehicle images has significant performance improvement over that of the corresponding original vehicle images.

The rest of this chapter is organized as follows. Section 4.1 describes our audio visual vehicle (AVV) dataset for moving vehicles. Section 4.2 presents a brief survey related to the MTP approach. Section 0 shows the MTP generation procedure. Section 4.4 describes the multimodal data alignment using the MTP. Section 4.5 describes the visual image reconstruction algorithm. Section 4.6 presents a physics-based LDV signal enhancement algorithm. Experimental results and conclusions are presented in Section 4.7 and Section **Error! Reference source not found.**, respectively.

4.1 Audio Visual Dataset

Using the designed multimodal system, we collected and built our own dataset of long-range moving vehicle classification. We will use those data to demonstrate the multimode sensing and processing framework throughout the rest of the thesis. There are two locations that the data were collected, one is at a local road and the other at a highway. The local road has number of occluded static objects, such as trees, parked vehicles, mailbox, and etc. This situation is very common in an urban environment where a lot of parked vehicles and trees on the road side. Fortunately, the traffic of the passing vehicles in the local road is sparse; thus, the data collected can be labeled for training purpose. On the other hand, although there is no occlusion on the road side, the traffic passing vehicles in the highway is very dense. Thus, those data will be labeled for testing. The stand-off distance of the multimodal sensor platform at the local road is about 25-30 meters. The stand-off distances for the highway data collecting vary from 50 to 70 meters. Various camera zoom levels are used to obtain images of adequate spatial resolutions. The acoustic signals are collected using the mono sound track of a sound card at 22.5 KHz of 16bit with input directly obtained from the LDV output. The data were collected at different days from Monday through Sunday with various weather conditions: sunny, cloudy, rainy and windy. The audio listening position was consistent through all days and times. The shortest video clip is about 1 minute and the

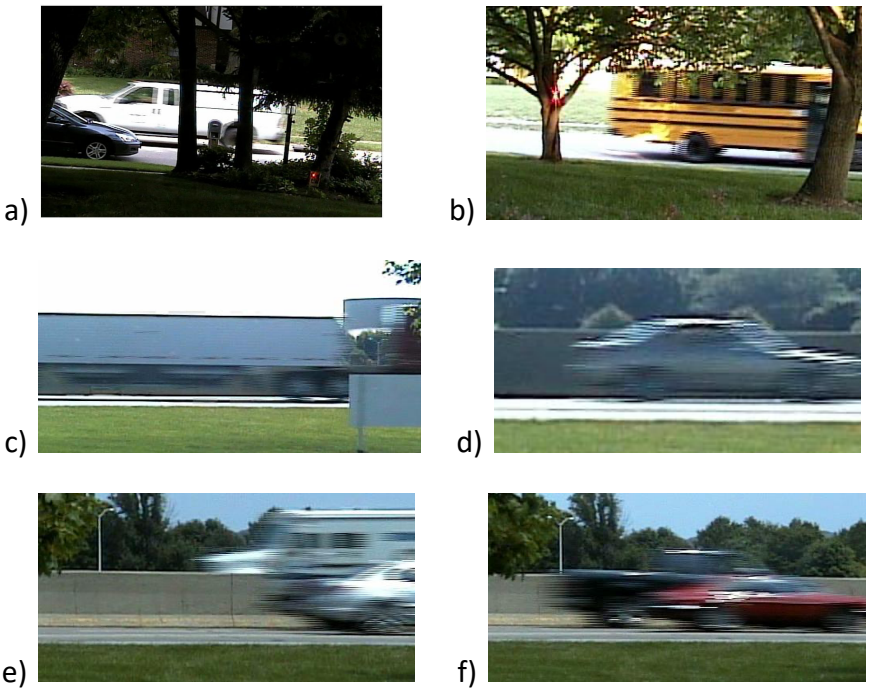


Figure 4.1 Challenges in vehicle detection and classification with long-range sensors.

longest one is about 3 hours. Each clip may contain zero, one, or many passing-by vehicles. There are about 3000 vehicle samples in total. Therefore, due to the large variations in recording durations, traffic volumes, camera setups, and scene locations, it is hard to manually label the data and process the raw clips. Figure 4.1 shows some challenging scenarios where moving vehicles are captured. The first two images (Figure 4.1 a and b) are captured at a local road: a white truck with motion blur and occlusions (a parked black car, trees, bushes, and a mailbox), and a bus with a large portion occluded. The mid two images (Figure 4.1 c and d) were collected at a highway: a very long truck only partially in the FOV, and a sedan moving in a high speed and showing obvious motion blur. Note there are always some cases two or more vehicles moving closely thus overlapped together in the image (Figure 4.1 e and f). Therefore, it is necessary to have

an efficient and effective data processing and representation technique for visualizing, searching and labeling the audio, visual and motion data of moving vehicles. Meanwhile, it is also desirable to remove perspective distortions, occlusions and motions blurs of vehicle images, and align them with corresponding acoustic signatures. This required a novel technical approach to vehicle reconstruction, labeling, and cross-sensor synchronization, which will be discussed in the next section.

4.2 A Brief Survey of Related Work

One of the earliest works using panoramic view images (PVIs) was route scene representation for robot navigation (Zheng & Tsuji, 1990); more recent works using the PVI concept to generate parallel-perspective panoramas for scene understanding can be found in (Seitz & Kim, 2003; Zheng, et al., 2006; Flora & Zheng, 2007). A 1D slit scanning approach was used to construct route panoramas when a camera is mounted on a moving vehicle. In these works, the resulted PVIs do not require inter-frame matching of video. The concept of the epipolar plane images (EPIs) was first introduced in (Bolles, et al., 1987), and it has been used to display features to trace the horizontal motion (Zheng & Wang, 2005). In (Flora & Zheng, 2006), they extract spatial-temporal information in the video volume and rectify the route panorama using two condensed image slices to record traces of horizontal and vertical scenes during the vehicle motion. They track the feature traces in the condensed image slices to remove jitter and adjust the local length of route panorama. In (Zhu, et al., 2000) both PVIs and EPIs are used for automatic traffic monitoring. However, vehicle shapes are not reconstructed therefore it will be hard to

classify vehicles based on the PVI images. All these papers only deal with video data, but in our work we also capture and process acoustic data using a LDV for better vehicle detection and classification. Works using both audio and video for surveillance can be found in (Dedeoglu, et al., 2008; Cristani, et al., 2007). In their approaches, the full video images are processed, which are sometimes computationally expensive but unnecessary. The synchronized labeling of the audio and video data for training classifiers could be very tedious. For example, moving vehicle detection does not really have to handle the change of the entire background or presence of other stationary objects that are irrelevant to the moving vehicles. So it would be ideal if we can only extract the audio and visual information of the moving vehicles and synchronize them in the time axis. Our MTP concept provides a way to only preserve the most important data that are both synchronized and normalized for extracting and labeling foreground moving objects.

4.3 Multimodal temporal panorama

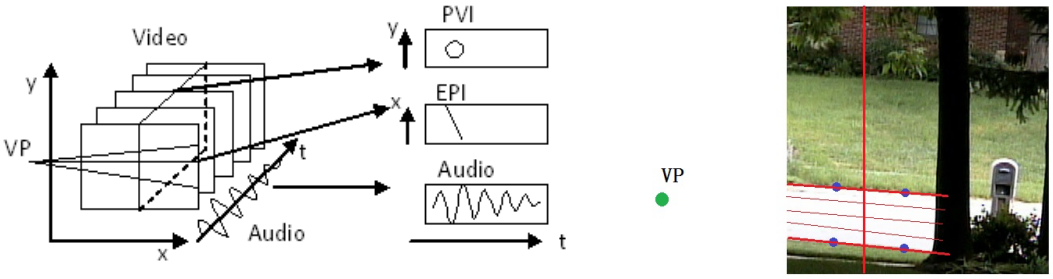


Figure 4.2 Temporal panorama generation and initial parameters selection

During data collection, both visual and audio data are captured simultaneously. Multimodal data including visual, motion, and audio information from moving vehicles are represented into a *multimodal temporal panorama* (MTP) as we first described in

(Wang, et al., 2011b). The MTP consists of three synchronized 2D spatial-temporal panoramas (Figure 4.2 left). The first panorama is the panoramic view image (PVI) concatenated from the same 1D vertical detection lines across all image frames (Zheng, et al., 1990; Zhu, et al., 2000). The least occluded line in the scene, particularly when there is a significant amount of occlusions such as trees, parked vehicles or others, is selected initially to detect any vehicles crosses the line (the vertical red line in Figure 4.2, right). In Figure 4.2, the right part shows a PTZ image overlaid with the least occluded vertical detection line (selected manually), and four points on the edges of the road to fit the two parallel road edges (two red vertical thick lines). Multiple "horizontal" epipolar lines that converge to the same vanishing point (in greed dot) are stored in a multiple-LUT. Only one of the epipolar lines is used that cuts through the middle of the vehicle detected in the vertical detection line. Using a single line approach ensures a consistent background subtraction result since there is little variation in consecutive background lines over time in the video sequence. The line can be reselected if the scene is changed or a new location is picked. The second panorama is the epipolar plane image (EPI) (Bolles, et al., 1987; Zhu, et al., 2000) that has same time axis as the PVI. The EPI is generated from concatenating 1D horizontal epipolar lines along the direction of a vehicle's motion. The purpose of this epipolar line is to track the motion of a vehicle on the road, after an initial target location on the road is selected. This location connects with the vanishing point of two parallel lines on the roadside to form an epipolar line. However, if the road is wide (i.e., a two way street, or a multi-lane road), a single fixed epipolar line may not be sufficient to trace the motion of a vehicle in various lanes/directions. A multi-look-up table (mLUT) is used to

store multiple epipolar lines that correspond with all possible moving paths of a vehicle. Once a vehicle is detected from the PVI, the right row index of the mLUT is selected for constructing the EPI. For both PVI and EPI, we do not require the whole body of a vehicle in the field of view. A partially viewed moving vehicle by the camera is sufficient. Last, a 1D audio wave scroll can be easily represented along with the PVI and the EPI in the same temporal axis. The first use of the audio information is to improve the robustness of vehicle detection using the PVI representation. The short time energy of a window of signals can distinguish a sounding target with silent background thus removing some false target detection from the PVI.

Figure 4.3 shows a segment of a MTP for a clip of multimodal vehicle collection on a local road. The duration of this clip is about 39 seconds. The synopsis is from the 3rd to 5th rows and the detection/reconstruction snapshots are on 1st, 2nd and 6th rows. First row shows the original frame snapshots that include vehicles. Second row shows the reconstructed vehicles. Third row shows the panoramic view image (PVI). Fourth row shows the epipolar plane image (EPI). Fifth row shows the audio wave scroll. Sixth row shows spectrograms and spectral energy plots of the detected vehicles. Here we want to note that due to the interlacing mechanism of the PTZ cameras, the images of moving vehicles are quite blurry. Therefore, we use fields as the unit for both PVI and EPI in the time direction instead of frames (consisting of even and odd fields). This reduces the spatial resolution of images in the vertical direction to half of the full resolution, but later in vehicle reconstruction, we will recover the original resolution.

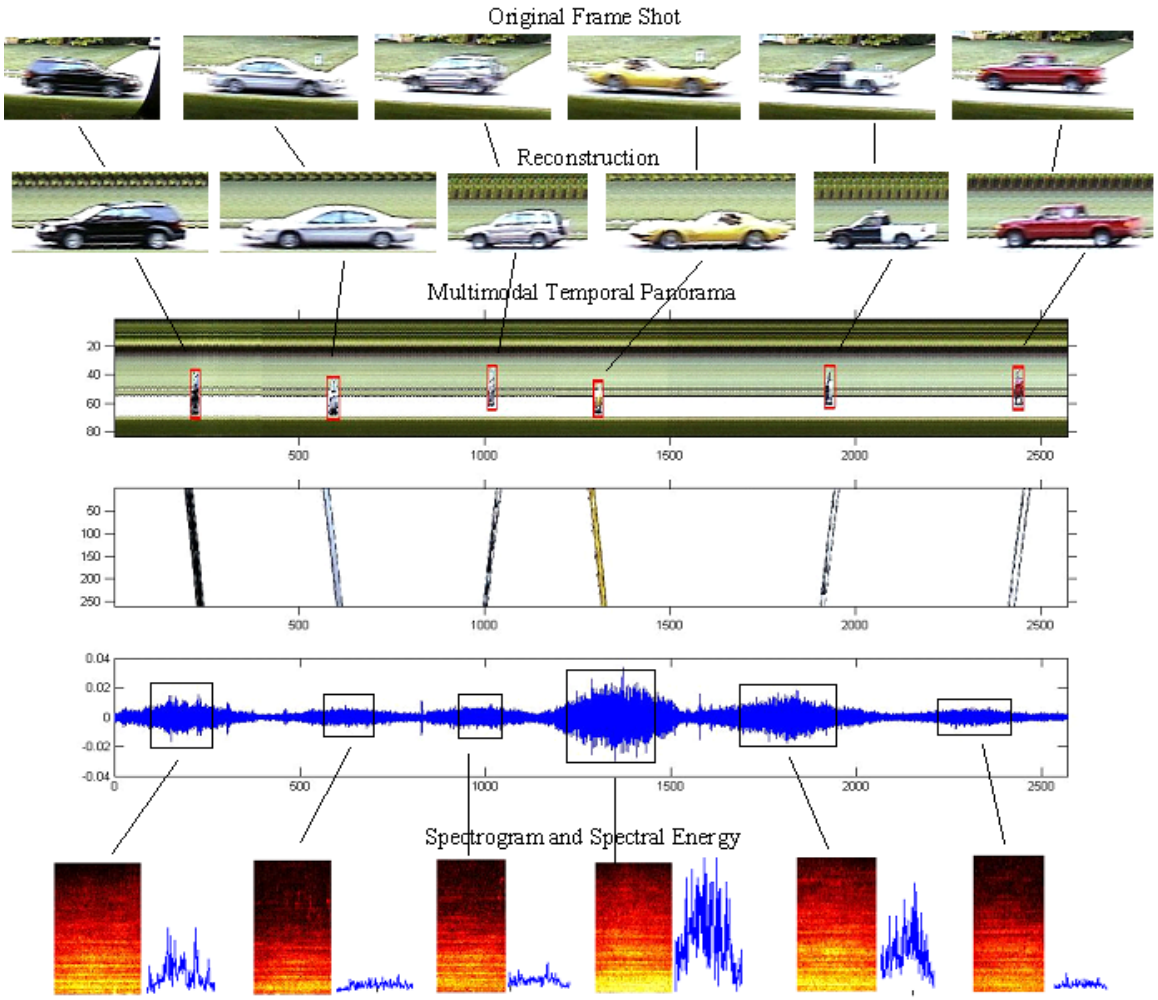


Figure 4.3 Multimodal temporal panorama on a local road

Figure 4.4 shows a segment of a MTP of multimodal vehicle collection on a highway. The duration of this clip is about 2 minutes. The synopsis is from the 3rd to 5th rows and the detection/reconstruction snapshots are on the 1st, 2nd and 6th rows. First row shows the original frame snapshots that include vehicles. Second row shows the reconstructed vehicles. Third row shows the panoramic view image (PVI). Fourth row shows the epipolar plane image (EPI). Fifth row shows the audio wave scroll. Sixth row shows audio energy plots of the detected vehicles. This scenario is more complicated than that of a

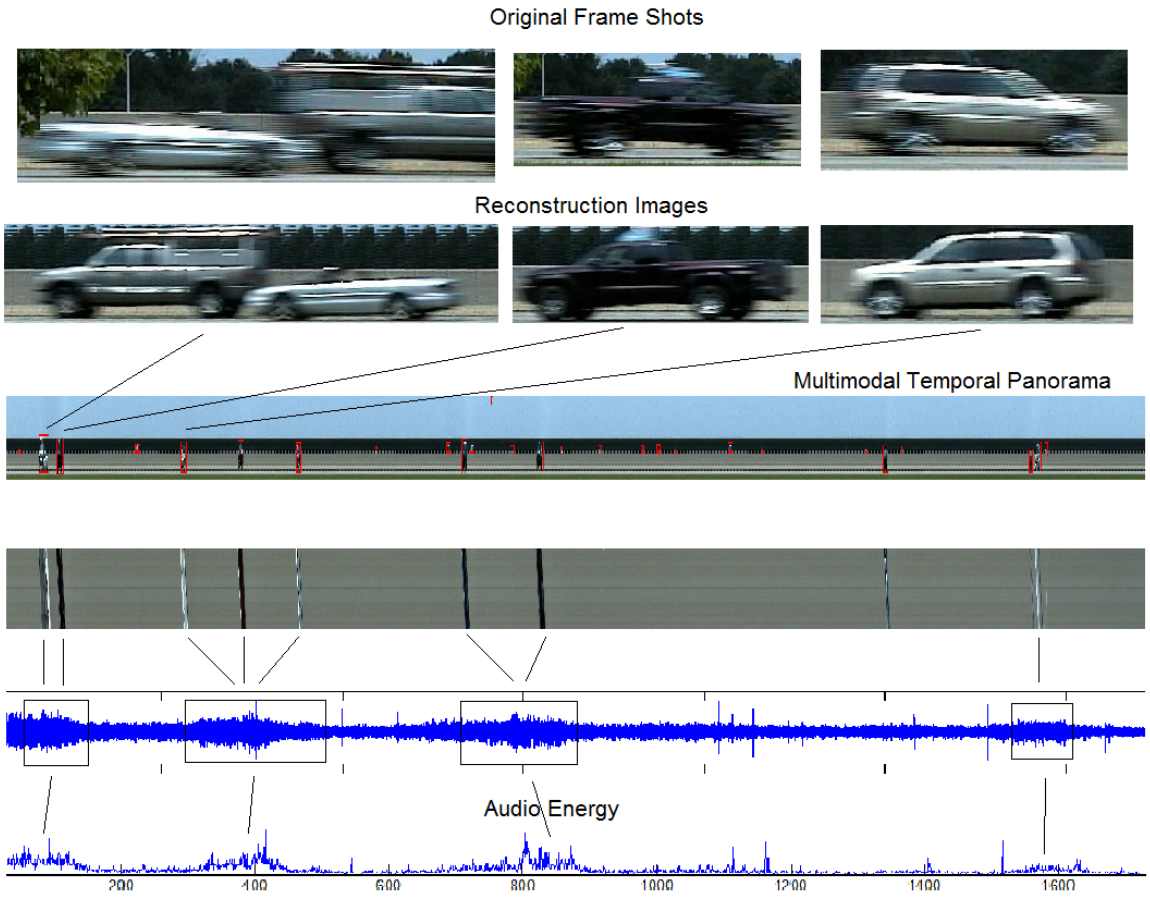


Figure 4.4 Multimodal temporal panorama on a highway road

local road where traffic is relatively sparse. In the first image shot, one vehicle is occluded by another visually; the sounds of the two moving vehicles next to each other with a very small headway are also mixed. For this case an image shot and a period of audio that contain these two vehicles are extracted together. Vehicles moving on the other side of the highway divider are also partially detected by the PVI. We detect them in the PVI but do not extract them for later classification based on the EPI and audio energy information.

The MTP facilitates the synchronization and integration of the information across the three modalities, both for automatic and interactive vehicle and traffic analysis, thus

providing more succinct and reliable information for tasks like moving vehicle detection and classification using visual, motion, and audio information.

4.4 Multimodal Data Alignment for Object Detection

Because our system allows us to collect multimodal data at different locations and selecting various detection zones, the visual detection and audio detection of vehicles may not be aligned. In other words, depending on the viewing angles of the camera and the directions of the moving vehicles, the system may hear the sound before or after it actually sees them. Also, noise from the background subtraction and ambient sounds may also cause invalid alignment. The three panoramas are first processed independently but simultaneously for object detection. Then results are combined and aligned to present the same objects in order to improve the detection rate.

4.4.1 Object Detection

During the generation of two spatio-temporal images, adaptive Gaussian mixture models (Stauffer & Grimson, 1999; Zivkovic, 2004; Zivkovic & van der Heijden, 2006) for background subtraction are applied for both PVI and EPI (Figure 4.5). In Figure 4.5, from top to bottom, the image shows the detection results (PDI, MDI, ADI) from the PVI, EPI and audio wave scroll after performing background subtraction. Note that only a small window of background containing a few lines is trained initially, and then new incoming lines are accumulated to update the model. It is much faster than performing the subtraction on the whole frames. Also the result is more consistent since there is little variation in consecutive background lines over time in the video sequence. Then,

morphological operations are applied to produce a panoramic detection image (PDI) and a motion detection image (MDI) corresponding to a PVI and its EPI in Figure 4.5, respectively. Note that there are still some false targets in both the PDI and MDI (in yellow oval shapes) due to other moving objects or background changes even after performing noise removal. However, along with audio information in the ADI, only objects with good audio signals are marked as vehicles (in red rectangle shapes).

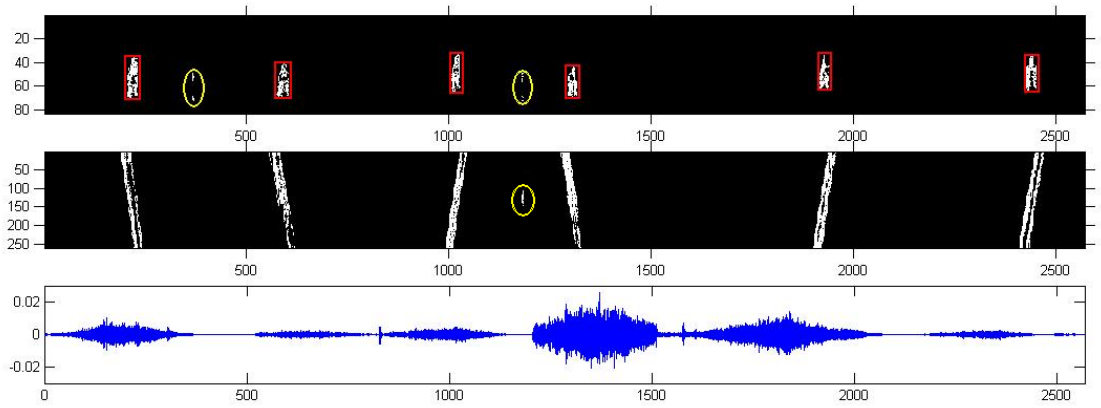


Figure 4.5 The processing results of the multimodal temporal panorama in Error! Reference source not found.

The process is performed online for every new frame, so the detection is done in real time. The PVI, which presents visual appearance of moving objects, is used for vehicle detection. In order to retrieve a vehicle from original frame shot, the center of the object region, which indicates the time frame in the original video, is used. **Error! Reference source not found.** shows original frame shots that have vehicles inside the field of view. The top and bottom boundaries of the bounding box of the vehicle can be easily determined from the PVI since the PVI has the same vertical coordinates as the original images. The determination of the left and right boundaries is not that straightforward.

So, the EPI is used for acquiring column pixel locations of the vehicle in the horizontal direction. The main purpose of the EPI is for estimating moving direction and speed of a vehicle. An increasing locus in the EPI indicates a vehicle move from left to right in the scene, and a decreasing locus means the opposite direction of a moving vehicle. The slope value of the locus indicates the speed of a vehicle. There are two possibilities for the slope shape, a straight line or a curved line. The straight line shows a vehicle move at a constant speed. It can be represented using a first degree polynomial function. If a vehicle accelerates or decelerates when it crosses the checkpoint, a curved line may be shown in the EPI. Usually it can be represented using a second degree polynomial.

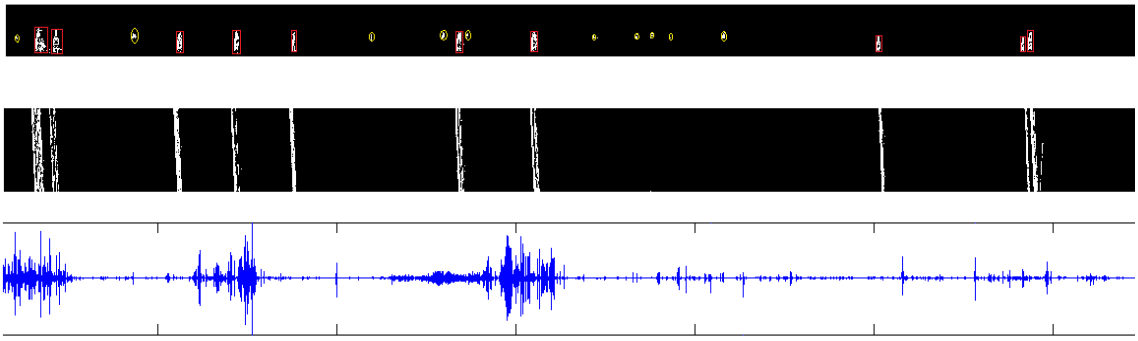


Figure 4.6 The processing results of the multimodal temporal panorama in Error! Reference source not found.

Using the same procedures for **Error! Reference source not found.**, Figure 4.6 shows the processing results for the MTP data in **Error! Reference source not found.**, where yellow ovals show the false target detection and the red boxes show the objects detected from the validation using both the EPI and the filtered audio wave.

In order to obtain good acoustic signature of an object, background noise for the scene is learned initially. Then a Wiener filter based two step noise reduction technique (Plapous,

et al., 2006) is applied to filter out noise that has corrupted an acoustic signal. The output *s'(t)* is defined as:

$$s'(t) = g(t) * [s(t) + n(t)] \tag{4.1}$$

where *s(t)* is the original signal to be estimated, *n(t)* is the noise, and *g(t)* is the Wiener filter's impulse response. In the snapshot layer of the MTP, snapshots of the spectrogram and the corresponding spectral energy are displayed for vehicles once they are detected in the PVI. This provides the capability to view the acoustic signature of a possible target.

4.4.2 Data Alignment

Next, we present a systematic way to align the multimodal data using the multimodal temporal panorama. Let I_i^D denote intensity map of a vehicle whose center body is detected at the time *i* in the appearance panorama *D*. Let I_i^M denote the intensity map displayed at the time *i* in the motion panorama *M*. We want to select a correct range *(j-m, j+m)* in an audio clip that corresponds to the detected vehicle, as:

$$\underset{j}{\operatorname{argmax}} \frac{1}{N} \left(\sum I_i^D + \sum I_i^M + \sum_{j-m}^{j+m} A\right) \tag{4.2}$$

where *j* is the center of the audio clip and *m* is the half-duration of the audio signal of a vehicle. *N* is the normalization factor, and *A* is the energy of the audio signals. Unlike human speech signals, the sound of a vehicle is much consistent during a period of time, so usually an audio clip of 5-10 seconds (for 2*m*) can describe the signature of a vehicle sufficiently. There are three main terms in this MTP base on Eq. (1): visual detection-$\sum I_i^D$, motion detection-$\sum I_i^M$, and audio detection-$\sum_{j-m}^{j+m} A$. A constraint that integrates these

three terms will be adjusted according to a specified task. For moving vehicle detection and classification, the motion of a vehicle has to be detected, together with either strong visual detection or strong sound detection. Here the either-or operation is used in case the vehicle could be very silent (such as an electric car). Thus, a constraint Ψ in subject to the Eq. 4.2 is set as:

$$\Psi = (\sum I_i^M > \tau)\left((\sum I_i^D > \tau)\ or\ \left(\sum_{j-m}^{j+m} A > \varphi\right)\right) = 1 \qquad (4.3)$$

where τ, φ are thresholds to penalize the visual background noise and the ambient sound. Note that the detection results do not rely on restricted thresholds. Indeed, a clean background subtraction with a filtered audio signal can guarantee a good detection results using very small τ, φ.

Error! Reference source not found. to Figure 4.5 show the detected objects in red rectangle boxes and their aligned audio clips in black boxes. For the first object (vehicle), the detection time of the visual appearance and the audio signals are different; however, they can be aligned by finding an audio region that yields the highest total energy with respect to the visual detecting region. In addition, false targets in the appearance panorama can be removed if there is no motion presented to indicate a moving vehicle. In other words, if there is a vehicle detected in all the three panoramas (appearance, motion and audio), the result should be 1 in Eq. (4.3); otherwise, the result should be 0. Note that the constraint is task dependent; and we assume a moving vehicle could be detected at both video and audio. It is definitely possible to hear the sound of a moving

vehicle without actually seeing it, then the constraint needs to be redesigned to fit in this situation

4.5 Reconstruction Algorithm

Vehicle reconstruction is necessary since the vehicles may be occluded by other stationary objects, such as bushes, trees, parked vehicles or others. Motion blur can also be removed after reconstruction. In addition, reconstructed vehicles all have the same views, which should improve the recognition and classification performance while keeping classifiers simple.

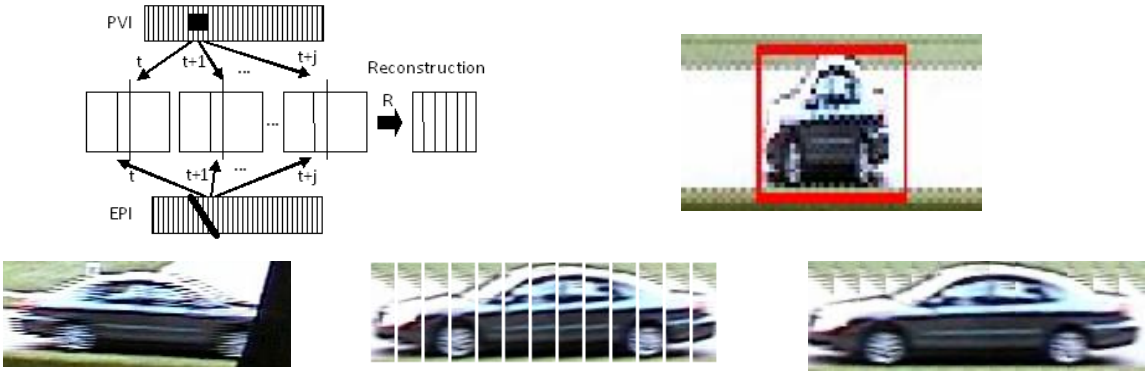


Figure 4.7 Vehicle image reconstruction procedure

The general idea of reconstruction is demonstrated in (Figure 4.7 top left). In Figure 4.7, the reconstruction procedure is illustrated on the top left. The top right shows an example of detected object from the PVI having even-odd field pair placed sequentially. The bottom shows the original zoomed image with motion blur and occlusion (left), un-rectified image pieces of a sedan with even-odd field aligned (middle) and its reconstructed result (right). Each vertical line in the detected region in the PVI indicates

a particular time frame I_t in the original video. The slope *m* of the vehicle's locus at the corresponding time *t* in the EPI shows the relative speed v_t of a moving vehicle as:

$$\mathrm{v_t} = \mathrm{m} = \frac{\partial \mathrm{x}}{\partial \mathrm{t}} \tag{4.4}$$

In other words, it is equal to the number of pixels in the motion direction in the original image that need to be extracted. The sign of the slope indicates the direction the vehicle moves to. So, if the vehicle moves from left to right, the image piece to the left of the vertical detection line (here defined as referenced line *rl*,) is extracted. If the vehicle moves from right to left, the image piece to the right of the *rl* is used. This is because the concatenation of PVI is in the left-to-right (or time increasing order). Then the image slice S_t at time *t* is:

$$\mathrm{S_t} = \mathrm{I_t^J},\ \mathrm{J} = \{\mathrm{j}|\mathrm{j} \in (\mathrm{rl}, \mathrm{rl} + \mathrm{v_t})\} \tag{4.5}$$

where *J* is the number of columns in the original time frame need to be selected. If the number is not an integer, then interpolation between two consecutive frames is applied. Although, the camera does not need to be perpendicular to the moving path of the vehicle, the segmented image pieces cannot be horizontal aligned smoothly if there is a rolling angle of the camera. An affine transformation is used to rectify those image pieces:

$$\mathrm{S_t} \mapsto \mathrm{A_\gamma S_t} + \mathrm{b} \tag{4.6}$$

where A_γ is the rotation matrix has rolling angle γ, and *b* is translation vector. If the true rolling angle is not known in advance, it can still be calculate from the initial image shot as:

$$\gamma = \tan^{-1} \frac{Ep_y - Vp_y}{Ep_x - Vp_x} \quad (4.7)$$

where (Ep_x, Ep_y) is the intersection point of the referenced vertical line and selected epipolar line; and (Vp_x, Vp_y) is the vanishing point of any two parallel lines showing the roads structure. Then the reconstructed image I_R for a vehicle is the integration of all image pieces from starting time t_s to finishing time t_f when the vehicle is observed through the reference vertical detection line:

$$I_R = \bigcup_{i=t_s}^{t_f} A_\gamma I_t^J \quad (4.8)$$

The first image shot of in **Error! Reference source not found.** shows two vehicles moving closely next to each other, so the reconstruction result contains both of them, overlapped. The motion slopes of those two are mixed together so only the best fitting line is selected to estimate the speed. So we can handle dense traffic, however we assume two or more vehicles moving at least at the same direction with a similar speed in order to make the reconstruction work.

The motion blur is mostly caused by the interlacing of the camera. Similarly, by knowing the speed of the vehicle, we can accurately align the even and odd fields of the image pieces into a single image piece at the frame *i*, thus significantly reducing the image blur, and restoring the original image resolution in the vertical direction. Figure 4.7 bottom shows an example before and after motion blur removal.

The reconstruction is based on the detection panorama and the motion panorama: the detection panorama shows the current position of the target pass across the vertical

detection line, and the motion panorama shows how many pixels need to be sliced from the corresponding original image shot. So it is the combination of image patches sliced from the original image at consecutive frames based on the relative speed. It is irrelevant to the distance and the camera since this is basically an image alignment problem, assuming the slices are of a planar surface of the car. We tested on data at different distances and different zoom levels, the shape of vehicles can mostly be reconstructed. Although a little shutter effect is still left, most conventional feature descriptors can be applied to distinguish the shapes of vehicles on the reconstructed images.

4.6 Audio Enhancement for LDV Signals

Similar to the vehicle image reconstruction, the purpose of audio enhancement is to make acoustic signals collected from the LDV having similar characteristics to those from typical acoustic sensors (i.e., microphones) and to make the extracted features more distinctive for target classification. Generally speaking, most microphones in use have very flat frequency responses whereas the frequency responses of LDV vary due to the surface vibration. For LDV voice acquisition, particularly the acoustic events occur from a large distance to the sensor, finding the right vibration surfaces close to the acoustic sources (humans, vehicles, etc.) is very important. Vibration measurements are made at the point where the laser beam strikes the target (surface) under the vibration caused by a voice source. Usually, the stricken targets have the structure of *plates.* Such targets include walls, doors, metal boxes, traffic signs, building pillars, containers, and so on (Zhu, et al, 2004).

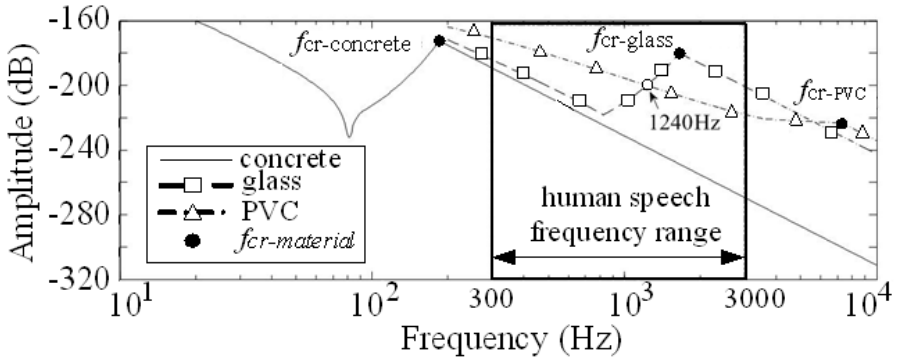


Figure 4.8 Vibration amplitudes on the top layers of concrete, glass and PVC

We have analyzed the vibration characteristics of several typical surfaces with different materials and structures are explored through both simulations and real-sensor experiments (Li, et al., 2010). Based on their responses to the frequencies in the range of human voice, the targets are classified into three categories by the number of fluctuations (zero, one or two) in their vibration returns in the range of speech (Figure 4.8). Both short and long range LDV voice detection experiments with these three kinds of targets verified our conclusion that, the acquired signals from the targets in the second and the third categories, like glass plates and paper boxes, give better performances and are recommended for LDV listening. Furthermore, the characteristic curves of frequency responses of these targets, can not only be used to make a better selection of appropriate surfaces for LDV voice detection, but also have the potential to be utilized for both signal enhancement and signal interpretation for the signals captured by the LDV off these targets. In this research work, we enhance the acoustic signal based on the experimental model.

4.7 Experimental Results

The reconstructed image of a vehicle not only indicates the detection of a vehicle at a time, but also provides a clean and complete visual appearance of the vehicle for data labeling, facilitating better vehicle classification. Here, we would like to first analyze the accuracy of reconstruction results, and then show the classification performance using reconstructed results against non-reconstructed images. Although the physic based acoustic signal enhancement is not the main contribution in this thesis, we still show a simple example of LDV signal enhancement.

4.7.1 Reconstruction Error Analysis

To show the accuracy of reconstructed image results, we perform error analysis under two cases depending on whether the true sizes of vehicles are known or not. For the first case, giving the true length *L* and the true height *H* of a vehicle, the relative errors of a vehicle in the length ε_L and in the height ε_L are:

$$\varepsilon_{\mathrm{L}} = \frac{|\mathrm{L}-\mathrm{L}'|}{\mathrm{L}}, \mathrm{L}' = \frac{\mathrm{I_L D_m}}{\mathrm{f_L}} \tag{4.9}$$

$$\varepsilon_{\mathrm{H}} = \frac{|\mathrm{H}-\mathrm{H}'|}{\mathrm{H}}, \mathrm{H}' = \frac{\mathrm{I_H D_m}}{\mathrm{f_H}} \tag{4.10}$$

where L' and H' are the length and the height in reconstructed result, respectively. I_L and I_H are the width and the height of the reconstructed vehicle image in pixels. f_L is the focal length in horizontal direction, and f_H is the focal length in vertical direction. Dm is the distance of a vehicle at the *mth* lane. We also perform a theoretical error analysis in order to compare with the actual errors calculated with Eq. (4.9) and Eq. (4.10). The theoretical relative errors in length ε'_L and in height ε'_L are:

$$\varepsilon'_{\mathrm{L}} = |\frac{\delta \mathrm{L}}{\mathrm{L}}|, \delta \mathrm{L} = \frac{\mathrm{D_m}}{\mathrm{f_L}} \delta \mathrm{I_L} \tag{4.11}$$

$$\varepsilon'_{\mathrm{H}} = |\frac{\delta \mathrm{H}}{\mathrm{H}}|, \delta \mathrm{H} = \frac{\mathrm{D_m}}{\mathrm{f_H}} \delta \mathrm{I_H} \tag{4.12}$$

where δI_H and δI_L are the measurement errors in the height and length directions of the image of a vehicle (in pixels).

If a vehicle's size is not known, we manually measure the length L'' and the height H'' of the vehicle in the original image corresponding to the reconstructed image at time t_m, where t_m is the time the vehicle half way passes through the detection line. Note that the vehicle may be partially occluded at the front or the rear part, or cannot be fully displayed in individual image frames. Therefore we combine the image frames at time t_s or t_f that have the vehicle partially displayed so that the correct length and height can be measured. Here t_s and t_f are the starting and finishing time the vehicle is detected. Then, the calculation of the relative errors for the unknown vehicle is just a matter of substituting *L"* and *H"* for *L* and *H* in Eqs. (4.9) and (4.10), respectively.

In our experiment, we had three vehicles with their known sizes provided by the manufacturers. They are Nissan Altima, Honda Accord, and Honda Pilot, each passing through the check point for 10 times on a two-way road about 24.8 and 26.8 meters to the camera. The focal length of the camera is 10.5 mm under 15x zoom. Some of the reconstruction results as well as their corresponding frame shots are shown in Figure 4.9. The top row shows the original image shots (zoomed and cropped), and the bottom row shows the corresponding reconstructed results. The actual relative error and theoretical relative error results are shown in Table 4.1. The theoretical errors are obtained by

assuming the measurement errors in the height and length directions of the image of a vehicle (δI_H and δI_L, in Eqs. 4.11 and 4.12) are both one pixel. The actual reconstruction errors are comparable to the corresponding theoretical errors, and the average reconstruction error in both length and height is about 4%. More reconstructed image results from the dataset are shown in Appendix C.

Figure 4.9 Sample reconstruction results for three vehicles: Nissan Altima (left), Honda Accord (middle), and Honda Pilot (right).

Table 4.1 Reconstruction error analysis for vehicles of known type

Type	Nissan Altima	Honda Accord	Honda Pilot	Total Avg. Err.
True L(mm)	4661	4811	4849	-
Act. err. in L	3.87%	5.31%	3.86%	4.34%
Theo. err. in L	3.87%	5.14%	3.70%	4.24%
True H (mm)	1420	1445	1847	-
Act. err in H	4.64%	5.37%	1.68%	3.90%
Theo. err in H	4.46%	5.28%	1.29%	3.70%

4.7.2 Classification on Reconstructed Results

For showing the effectiveness of vehicle reconstruction and background removal, we apply the HOG feature extraction on three sets of images of the same vehicles: original

raw images, reconstructed images without background removal, and reconstructed images with background removed. We used 667 samples, 400 are used for training and 267 for testing. The vehicles are labeled into four categories: sedans, vans, pickup trucks and buses. There are many variations in each category. For example, sedans contain sports cars and economic 2 door or 4 door cars some with fastback or hatchback; vans include mini vans, regular size vans and long size vans, note that SUVs are categorized into vans as well; pickup trucks some may have wagons or trailers at rear parts; and buses include both school buses and transportation buses. The original images are directly cut out from the original video frames that best correspond to the reconstructed results. Note that the original images may include partial occlusions, various side views and motion blurs. Table 4.2 shows the comparison in three confusion matrices on the same testing data, where the rows indicate ground truth and columns are the estimations. The training size, testing size and training parameters are all the same for the three sets of data. Applying the same classifiers, the reconstruction without background removal improves the performance by 15.73%, and reconstruction plus background removal improves the performance by more than 18.10%. Therefore, from this point on, the HOG features are extracted only from reconstructed images with background removal.

Table 4.2 Performance improvement with reconstruction & background removal (S-Sedans, V-Vans, T-pickup Trucks, B-Buses).

Original images				
Accuracy: 54.31%				
	S	T	V	B
S	77	3	26	2
T	10	7	11	0
V	49	6	56	2
B	7	3	3	5

Reconstruction only				
Accuracy: 70.04%				
	S	T	V	B
S	81	6	20	1
T	3	23	2	0
V	25	12	72	4
B	4	0	3	11

Reconst - background				
Accuracy: 72.41%				
	S	T	V	B
S	83	3	20	2
T	4	22	2	0
V	26	4	80	3
B	3	1	3	11

4.7.3 Results of Audio Enhancement

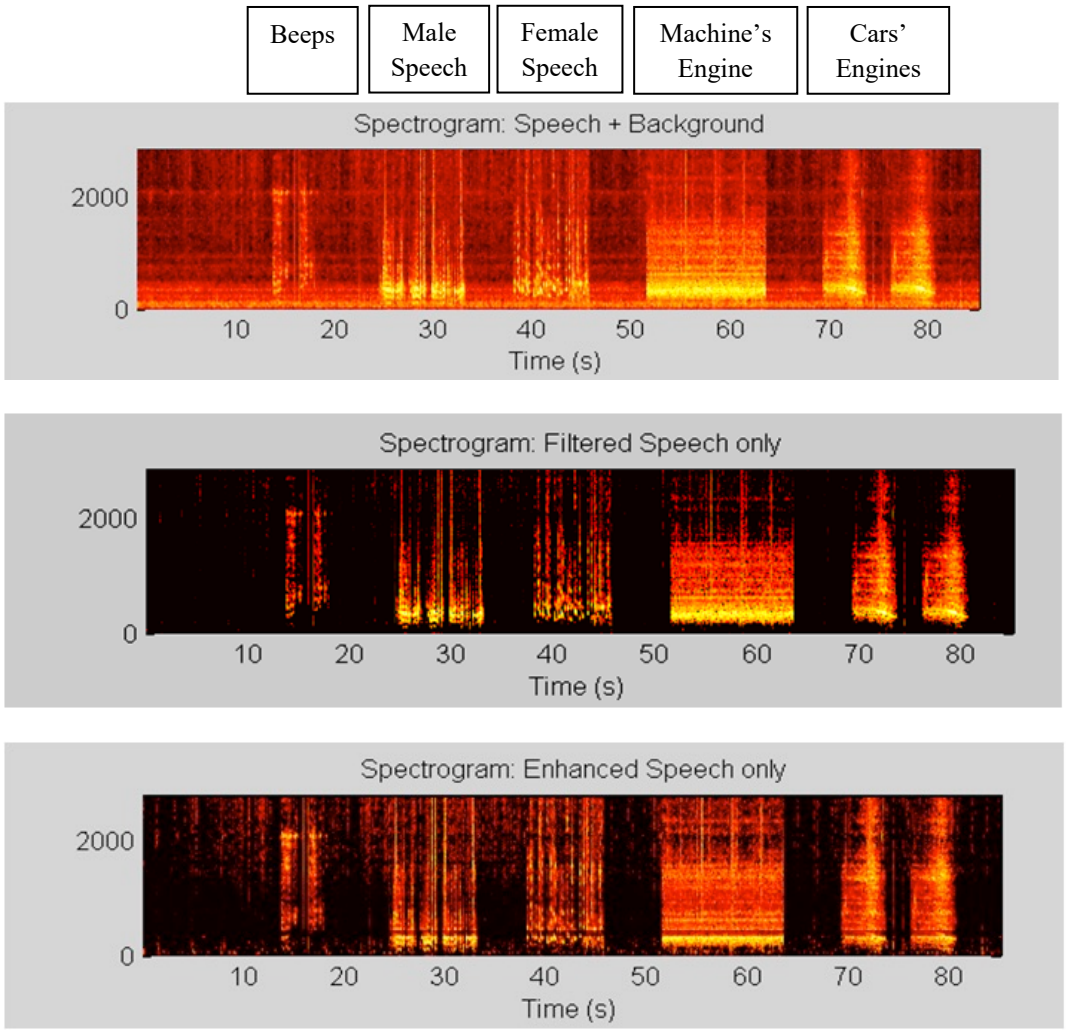


Figure 4.10 Spectrograms of original sound (top), filtered sound (middle), and enhanced sound (bottom)

In Figure 4.10 we show an example of audio speech with and without enhancement. The top spectrogram shows the original audio data include speech with a machine's engine sound and two cars' engine sounds. Usually, a band pass filter or Weiner filter will remove low frequency noises shown in middle spectrogram. This assumes the frequency responses are flat for signals at all frequency range (true if using microphone). However, LDV may respond strongly at one frequency range but weakly at another frequency range for a surface. We need apply some enhancement to recover some information at some frequency range in order make distinct of two classes, say machine engine with car engine. The bottom spectrum shows the enhanced audio data that makes the last columns (machine engine and car engines) differently.

4.8 Concluding Remarks

In this paper, we first describe the new audio-visual dataset acquired for moving vehicle detection and classification. Noisy multimodal data are represented efficiently in a multimodal temporal panorama interface for automatic moving target detection and fast vehicle labeling. A visual image reconstruction technique is provided to improve vehicle classification. A physics-based audio enhance technique is used to remove background noises while keep some low frequency signatures of the vehicles.

Chapter 5

5 Multimodal Feature Extraction

Thanks to the MTP approach, various types for visual features and audio features can be applied for the objects. With the reconstructed visual images, both scale and metric features as well as view and scale invariant features can be used. The visual features include aspect ratio and size (ARS), histograms of oriented gradients (HOGs), shape profiles (SP), representing simple global scale features, statistical features, and global structure features, respectively. The audio features include short time energy (STE), spectral energy, entropy, flux and centroid feature, and Mel-frequency cepstral coefficients (MFCCs), which are grouped into three types: temporal features (STEs), spectral features (SPECs) and perceptual features (PERCs).

In this chapter, we start with a brief overview of feature extraction, particularly for visual and audio features, in Section 5.1. Then we describe visual feature extraction in Section 5.2, and audio feature extraction in Section 5.3. A brief explanation of how multimodal feature combined and synchronized is presented in Section 5.4. Some sample results are shown in Section 5.5. Conclusions are provided in Section **Error! Reference source not found.**.

5.1 A Brief Overview of Feature Extraction

An image feature set should represent the most relevant information for object detection and classification, meanwhile providing invariance to changes in illumination, differences

in viewpoints and shifts and size changes in object contours. Many local image features have been proposed, such as points (Mikolajczyk & Schmid, 2002), blobs (Lowe, 2001), intensities (Vidal-Naquet & Ullman, 2003), gradients (Ronfar, et al., 2002; Mikolajczyk et al., 2004, Dalal & Triggs, 2005), color, texture, or combinations of several or all of these. Despite the diversity, they could be roughly divided into two broad categories: (1) sparse feature representations based on points, image fragments or parts; and (2) dense feature representations using image intensities or gradients. Sparse feature representations are based on local features of local image regions that can be selected using key point detectors or part detectors. A dense feature representation extracts image features over an entire image or detection window, which are collected into a high-dimensional descriptor vector that can be used for discriminative image classification. Using image features as the basis, video features provide additional temporal information via either optical flow or motion tracking techniques.

The audio features can be categorized into three groups: time-series features, spectral features, and perceptual features. The time-series features represent audio samples in their raw waveforms, and may be enhanced by some filters, such as zero crossing rates or short time energy (Lu, et al, 2002). The spectral features represent spectral moments and flatness, such as spectral centroid, spectral roll-off, spectral flux and linear coefficients (Rabiner and Juang, 1993). The perceptual features represent the spectral variation and sharpness, such as Mel-frequency cepstral coefficients (MFCCs), or delta MFCCs (Knox & Mirghafori, 2007). A complete description of some typical and commonly used audio features is presented in (Teodoridi & Koutroumbas, 2008). However, there is

no direct proof showing that one performs better than others because feature generation is very much task and data dependent. Here, we select some common and reliable features for testing our feature modality selection algorithm in Chapter 5.

5.2 Visual Feature Extraction

The visual features are extracted from the reconstructed image results. The objective of reconstruction is to make vehicles' visual images invariant to perspective views and distances. Also, the results have occlusions and motion blur removed. Therefore, both metric features as well as statistical features can be used more effectively. The first feature that can be used is simply the aspect ratio and size (ARS) of the vehicle, as $f_{ARS} = [w, h, w/h]$, where w is the width and h is the height. It can classify vehicles into various sizes. Note that even though moving vehicles can be captured at different distances or camera zoom levels, the reconstructed image results are invariant in size since the distances are measured via the PTZ stereo and the images can be normalized.

The other visual feature is the shape profile (SP), which is a curve that indicates the top boundary of a vehicle, a strong indicator of the vehicle's type. To create the SP, we first apply the background subtraction on image pieces of the reconstructed vehicle image to obtain a clean shape of a vehicle. Only the top half of the images is used since only the top boundary contains significant differences among different types of vehicles, and the bottom part is harder to segment from the background due to shadows and motion blurs of the wheels. Then, the top boundary curves of all the vehicle images are sampled into the same number of bins with each bin B_i presents the average of height of the current

shape boundary, and to form a feature vector f_{SP} of the same dimension after normalization as:

$$f_{SP} = \frac{1}{max\ B}[B_{1,}B_{2,} \dots, B_N] \quad (5.1)$$

where *N* is the number of bins for the *SP*. Note this normalization loses the size information. Note that this normalization loses the size information, but it has been captured by the aspect-ratio and size feature.

Histograms of oriented gradients (HOGs) (Dalal and Triggs, 2005) are a statistical feature that preserves some texture and local structure. It counts occurrences of gradient orientation in localized dense grid cells uniformly, thus, forming a feature vector of histogram H as $f_{HOG} = [H_1, H_2, ..., H_M]$, where M is the number of bins for the HOGs. Since it uses local contrast normalization, it is invariant to illumination changes, thus, it is good at people detection as well as vehicle detection (Mao, et al., 2010). We also extract HOGs for both reconstructed vehicle images with and without background removal for comparison of classification performance. By representing multimodal data in multimodal temporal panoramas (MTPs), static HOG features can be effectively applied to the reconstructed images of vehicles. The number of cells and histograms used depend on the sizes (or resolutions) of the reconstructed images. For wide-area surveillance, the monitored moving object may be small after reconstruction, so a small number of cells could be good enough to characterize the visual signature of the object. However, if we zoom in the camera to get fine details of the object, small number of cells could not distinguish various types of objects significantly. For our current experiments, we use 3x6

cells which three rows present vehicles top, middle and bottom parts and six columns present vehicles front, middle body and read parts.

5.3 Audio Feature Extraction

In general, audio features can be categorized into three groups: time-series features, spectral features and perceptual features. The time-series features represent audio samples in their raw waveforms. Short time energy (STE) is used to calculate the energy over a time (Lu, et al., 2002). It is usually good at distinguishing a vehicle's sound with a silent background. Since the audio signals of a moving vehicle are much consistent over a short time period, overlapped windows in a short period of time clip are used to calculate the sound energies of the detected corresponding object (vehicle) in the PVI. Then we form the STE feature vector using only their mean and standard deviation as: $f_{STE} = [\mu_{STE}, \sigma_{STE}]$.

In the second group, the spectral features (SPEC) represent spectral moments and flatness (Rabiner and Juang, 1993). Spectral energy, entropy, flux and centroid are composed together into a spectral feature vector $f_{SPEC}=[Eng, Ent, Flux, Cent]$. The spectral energy *Eng* calculates the energy of the power spectrum defined as:

$$Eng = \sum |F\{x(t)\}|^2 \tag{5.2}$$

where x(t) is the audio signal and F{} is the Fourier transform. The spectral entropy *Ent* measures the energy changes and defined as

$$Ent = -\sum |F\{x(t)\}| \frac{\log |F\{x(t)\}|}{E} \tag{5.3}$$

The spectral flux *Flux* measures how quickly the power spectrum of a signal is changing and defined as:

$$Flux = \sum(|F\{x(t)\}| - |F\{x(t-1)\}|)^2 \tag{5.4}$$

The spectral centroid *Cent* indicates the center of the spectrum defined as:

$$Cent = \frac{\sum w|F\{x(t)\}|}{\sum |F\{x(t)\}|} \tag{5.5}$$

where *w* is the weighted mean vector of the same dimension as the *F*.

In the third group, the perceptual features (PERC) represent the spectral variation and sharpness. Mel-frequency cepstral coefficients (MFCCs) (Zheng, et al., 2001) are commonly used to perceptually represent the frequency band responses of the human auditory system. The mel-frequency cepstrum (MFC) equally spaces the frequency band on the mel scale of *F{x(t)}*, and then transformed using the DCT after log of powers at each mel frequency. Then the coefficients of the results forms the perceptual feature $f_{PERC} = [\mu_{MFCC}, \sigma_{MFCC}]$, where μ_{MFCC}, σ_{MFCC} are the mean and the standard deviation vectors of all coefficients, respectively.

However, due to the noises from the LDV's electronic-optical effects and unforeseen environmental effects, the "foreground" acoustic signals (such as the sounds of human speeches, vehicle engines, etc.) may not stand out clearly from the "background" noises. A reliable background modeling technique should be employed that is distinct for different surface types but identical for the same type at various distances. A Gaussian Mixture Model (GMM) Φ is commonly used to model the feature distribution of signals

using a weighted summation of a Gaussian distribution *N*. The likelihood of a feature vector *x* is defined as:

$$\Phi(\mathrm{x}) = \sum_{k=1}^{K} \alpha_k N(x, \mu_k, \Sigma_\mathrm{k}) \tag{5.6}$$

where μ_k and Σ_k are mean and covariance matrix of *k*th Gaussian among *K* Gaussians, and α_k is a normalizing factor in range between 0 and 1. Due to the variations of the vibration properties from surface to surface, the GMM on each selected surface are constructed differently. Because the number of Gaussians *K* is different for each unique model, we have to use the right model for the right surface and evaluate the correctness. Also note that the GMM can model the feature distribution, however, it cannot present temporal dependencies of each component. In order to model the internal dynamic between the components distribution in temporal domain, we use a score-based aggregation technique for the GMM with more than one component (Wang, et al, 2010b).

5.4 Multimodal Feature Synchronization

In moving vehicle detection and classification, the proposed multimodal temporal panorama (MTP) approach can represent and align data from multimodalities in the same temporal domain. It significantly improves the synchronizing process. Then we only need to map various types of features that represent the same object. The feature level integration has the advantage of sharing the information from various modalities, whereas the decision level integration could treat each modality independently.

First, we define the synchronization process as: given a set of features for the modality *M1* as $F_{M1} = [f_{11}, \ldots, f_{1P}] \in R^P$, and another set of features for the modality *M2* as F_{M2} =

$[f_{21}, \ldots, f_{2Q}] \in R^{Q}$, we determine the mapping $G(f)$: $f_{1i} \leftrightarrow f_{2j}$, where $i \in P$ and $j \in Q$. With the help of MTP alignment and reconstruction, this synchronization process is made easier. As a matter of fact, it is just concatenation of different feature modalities into a large vector. However, many samples of audio features may be extracted from a period of audio clip that corresponds to the same object in the visual image. We simply take the average of all of them for representing the audio feature of the same object.

5.5 Sample Results

In our current experiments, we used the data acquired on a local, 2-way road, with 667 different vehicles in the dataset. Note that this is from the same source of data of our previous work (Wang and Zhu, 2012a), but the number of vehicle samples is almost tripled and more multimodal features are extracted and analyzed. (Previously we only used HOGs for visual and MFCCs for audio modalities.) Of the 667 vehicle samples, 400 are used for training and 267 for testing. All vehicles' visual images are reconstructed so that the vehicle image results are invariant to perspective views, and the occlusions and motion blurs are removed. In our experiments, the vehicles are labeled into four categories: sedans, vans, pickup trucks and buses. There are more variations in each category than the dataset we used in our previous work (Wang and Zhu, 2012a). For example, sedans contain both sport cars and economic 2 door or 4 door cars; some with fastback or hatchback. Vans include mini vans, regular size vans and long size vans; note that SUVs are categorized into vans as well. Pickup trucks some may have wagons or trailers at rear parts. Buses include both school buses and transportation buses.

The ARS feature includes height, length and length/height ratio of vehicles. For the HOG feature, each vehicle image is divided into 6x3 grids and each grid has 9 bins so that the result HOG feature vector for a vehicle image has 162 dimensions. The SP feature uses normalized 30 bins across the top profile of the vehicle. The STE feature consists of a mean and a standard deviation of a vehicle temporal energy. The spectral feature contains means of 4 different types of spectral features and their standard deviations. For the perceptual features, we use the first 15 coefficients of MFCCs and calculate their means and standard deviations into a feature vector of 30 dimensions. Examples of the four types and some sample features are shown in Figure 5.1. The first row shows the best original image shots based on the automatic detection. The second row shows the reconstruction results (with background). The third row shows the reconstructed images with top half background automatically removed. The fourth row shows the shape profiles. Base on observation, the four different types of vehicles can be easily distinguished by using the shape profiles. The fifth row shows the histograms of oriented gradients in dense grid cells that form the typical HOG feature vector. Audio features in term of wave form (each 3 seconds), spectral domain are cepstral domains are shown in row 6, 7, and 8 respectively.

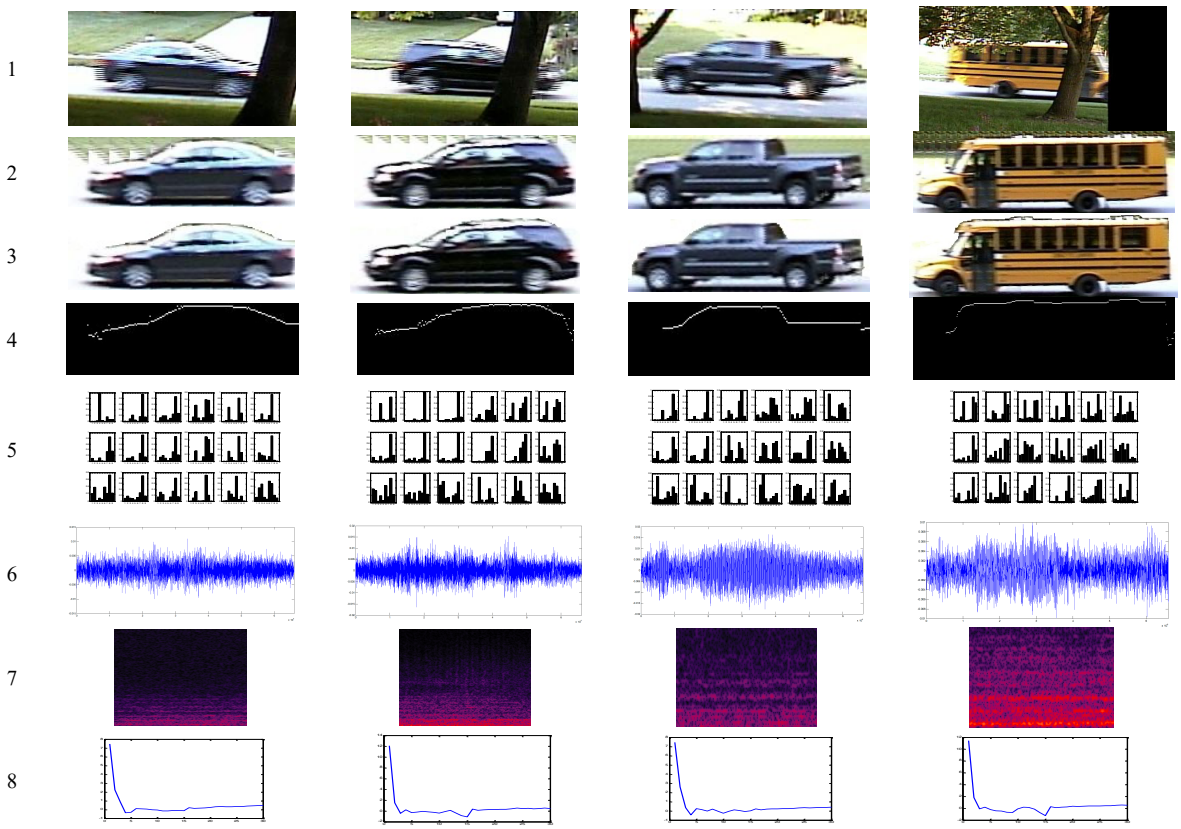


Figure 5.1 Samples of multimodal data of vehicles in four categories (sedan, van, truck and bus)

5.6 Concluding Remarks

In this chapter, both metric and statistical visual features are extracted from the reconstructed vehicles' images. The audio features are extracted based on three types: temporal features (STEs), spectral features (SPECs) and perceptual features (PERCs). The selected features are supposed to be representative for various types of information (shapes, sizes, sounds, etc), and we expect they can provide complementarities to each other.

Chapter 6

6 Multimodal Feature Selection and Learning

Multimodal feature selection and learning is particularly important for multimodal sensing where heterogeneous data are unavoidable, and we often cannot determine what the most appropriate features are, and how the feature of different modalities should be combined. In Chapter 4, we represent multimodal data into multimodal temporal panorama (MTP) that facilitates the synchronization and integration of the information across various modalities, thus providing more succinct and reliable information for tasks like moving vehicle detection. With multimodal sensing data, we are interested in various scenarios of multimodal classification, which involves the selection of multimodal features that we have discussed in Chapter 5 for given tasks. The issue of what to fuse has been addressed at two different levels: feature modality selection and feature vector reduction (Potamianos, et al., 2004). Feature modality selection refers to choosing different types of modalities, which could be different sensor sources or heterogeneous features extracted from a single sensor source. For example, a moving vehicle can be detected via a video camera or a microphone, or both. On the other hand, the fusion of features usually creates a large feature vector so that many feature reduction techniques are applied to overcome this problem. A lot of work has been done on feature vector reduction (Wall et al., 2003; Chetty & Wagner, 2006; Potamianos et al., 2001). But there is relatively little work on feature modality selection. Therefore, we mainly focus on feature modality selection.

The rest of chapter is organized as follows. Section 6.1 discusses some related work of feature modality selection and feature vector reduction. A multi-branch feature searching (MBFS) technique based on sequential forward selection algorithm is described in Section 6.2, also can be found in (Wang & Zhu, 2012b). Section 0 describes the boosting-based feature learning (BBFL) technique. Experimental results on feature selection and learning are shown in Section 6.4. Conclusions are provided in Section **Error! Reference source not found.**.

6.1 Related Work

The feature modality selection problem has often been considered as an optimization problem satisfying some conditions. For example, a moving vehicle can be easily detected and classified from the video analysis than the audio analysis if the vehicle's appearance can be observed clearly. However, if there is a large obstacle occludes most part of the vehicle, the audio analysis could be more handy. Oshman (1994), Debouk et al. (2002) and Jiang et al. (2003) mainly focus on sensor modality selection. In term of optimal feature modality selection, Wu et al. (2004) first find statistically independent modalities from raw features, then determine the optimal combination of individual modalities using a super-kernel. When all feature components were combined and treated as a one-vector representation, it suffers from the curse of dimensionality. On the other hand, the large number of modalities reduces the curse of dimensionality, but the inter-modality correlation increased. An optimal value of modality is selected to balance between the

curse of dimensionality and the inter-modality correlation. A summary of the approaches proposed in the above papers can be found in (Atrey et al., 2006).

The fusion of features that are obtained from different modalities usually result into a large feature vector, so that many feature reduction techniques are applied. Commonly used are principle component analysis (PCA), and linear discriminant analysis (LDA). PCA is used to project higher dimensional data into lower dimensional space while preserving as much information as possible. LDV is used for determining the linear combination of features, which is not only a reduced set of features but it is also used for classification. Wall et al. (2003) provide more details about these feature dimensionality reduction methods. Many researchers have used these methods for feature vector dimension reduction for the multimodal fusion, for example: Guironnet et al. (2005) used PCA for video classification, Chetty and Wagner (2006) utilized singular vector decomposition (SVD) for biometric person authentication, and Potamianos et al. (2001) adopted LDA for speech recognition.

From the available feature set, which modalities should be selected to accomplish a specified task? The utility of those modalities could be different given different tasks. As the optimal feature subset changes over time, how confidence the feature modality selected with which the task is accomplished, is an open problem for multimodal feature fusion and classification. Since most work have discussed on feature vector reduction but few of them discusses the reliable methods how feature modality selection, we will present a multi-branch feature searching technique in Section 6.2 and a boosting base

feature learning technique in Section 0, both for feature modality selection. Then in Section 6.4, we will provide some experimental results and comparisons.

6.2 Multi-Branch Feature Searching (MBFS)

Feature selection is a task dependent problem. Given two different tasks, the classification results may be different using the same features or feature combinations. We'd like to evaluate a large number of features and select only a few of representative features or feature combinations. Such problem can be formulated as: given a feature set $F = \{f_i | i=1,\ldots, N\}$, find a subset S_M with M < N, that maximizes an objective function $J(S)$,

$$S_M = \{f_{i1}, f_{i2}, \ldots, f_{iM}\} = \underset{M,iM}{\operatorname{argmax}} J\{f_i | i = 1,..N\} \tag{6.1}$$

The commonly used selection strategy is sequential forward selection (SFS) (Gheyas & Smith, 2010), which starts from the empty set and sequentially adds the feature that maximizes $J(S)$ in each step, then the process is repeated testing each remaining feature combinations with those previously preserved until all features have been evaluated. The problem of the SFS algorithm is that only a single best feature is selected at each round so that it has a tendency to become trapped in local maxima. To alleviate this problem, we design a *multi-branching feature searching* (MBFS) technique based on sequential forward selection algorithm which selects a number of good features at each round (level) that satisfy some maximal $J(S)$ above a threshold ω. First, let us define the following symbols:

N: the number of uni-modal features.

K: the number of levels of feature combinations, k=1: uni-modal, k=2: bi-modal, and so forth.

M_k: the number of selected features and/or feature subsets at the level k.

S_M: the M_th selected feature subset, where $M = M_1+M_2+\ldots+M_k$.

F': an available feature set, a subset of F.

F*: features in S_M.

$\omega(f)$: classification accuracy of a feature or feature combination, *f*

ε: a small tolerance value.

In the first level, a classifier is trained for each of the N uni-modal features and its classification accuracy is calculated. Then a subset S_{M1} with the top M_1 uni-modal features are selected whose classification accuracy drops from the best one is within a small percentage ε. Then in the second level, each of these M_1 features will be paired with the other un-selected features in the first round to generate multiple bi-modal features to train their classifiers. The same selection rule is used to select the top M_2 bi-modal features. This process continues to level K and therefore the selected feature subset S_M include M features or feature combinations, and $S_M = S_{M1} \cup S_{M2} \ldots \cup S_{MK}$. Last, the feature with the best accuracy among all levels of feature combinations in S_M is selected. This usually will be a multimodal feature, but it could be a unimodal or bimodal feature.

The algorithm is formulated as the following:

1. Start with the empty set $S_0 = \{\varnothing\}$, k=1;
2. *Let* $F'=\{f'_i | i=1,\ldots,N\}=F$;
3. Select the next subset of *k*-modal features S_{Mk} in the level *k*, by combining a feature f'_i in F' with every feature F^*_j in the subset $S_{M(k-1)}$, $j=1,\ldots,M_{k-1}$, s.t. $\omega(F^*_j+f'_i) \geq \omega_{max} - \varepsilon$ and $F^*_j+f'_i \notin S_M$ where ω_{max} is the accuracy of the best classifier in *kth* level;
4. Update $S_M = S_{M(k-1)} \cup S_{Mk}$;
5. $F'=F'-\{f'_i\}$, if $F'=\{\varnothing\}$, k=k+1, go to step 2, else go to step 3.

For both unimodal and multimodal features, the radial based support vector machines (SVMs) (Cortes and Vapnik, 1995) are used. For the multi-class problem, one-against-one technique is used by fitting all binary sub-classifiers and finding the correct class using a voting mechanism. To evaluate the classifier for a given feature or feature set, confusion matrix *C* is generated and its error $\varepsilon=1-trace(Diag(C)/sum(C))$ is calculated to indicate what percentage the true labels and expected labels are off the diagonal. As a result, only the best feature modality or the combination of feature modalities is selected.

6.3 Boosting Based Feature Learning (BBFL)

Boosting is a rather general approach for improving the performance of any weak classifiers. Here a weaker classifier is defined as any classifier that can achieve classification accuracy above 50%. Classification performance is boosted by combining many weak classifiers to produce a strong classifier. In the boosting literature, feature fusion is achieved by using the available features to create a new combination of these features. One example is the method of Kegl and Busa-Fekete (2009) which learns products of decision trees. Alternatively, Danielsson, et al. (2011) suggest the addition, to the boosted classifier, of logical (and, or) combinations of previously selected weak learners. Saberian and Vasconcelos (2012) derive more sophisticated combinations of weak learners for boosting feature selection and extraction. The resulting boosting algorithms grow a predictor by selecting among a pair of pre-defined operations, which could be sums and products or “ands” and “ors”, among others. However, their work cannot be applied to feature modality selection directly.

The basic idea of our boosting-based feature learning (BBFL) is to not only learn the weak classifiers given input training samples, but also learn the weak classifiers with the selected feature modalities. Then the “winner-takes-all” approach selects the best classifier of the corresponding feature modalities. Our method uses the exhaustive search that learns weak classifiers for all feature modalities and their combinations. In our experiments, we use decision trees as the weak classifiers.

The original AdaBoost works for binary classification problems. For multiclass problems, a meta-classifier is designed for general n-class problem. Two straightforward combination schemes are the one-again-all classifier and the one-against-one (or pairwise) classifier (Tax & Duin, 2002). With the one-against-all classifier, n classifiers are trained, each of which is able to distinguish one class from all of the others. At the end, the testing vector is assigned the class corresponding to that of the machine producing the largest positive score. The one-against-one classifier uses $\frac{(n)(n-1)}{2}$ binary classifiers to separate each class from each other class. A voting scheme is then used at the end to determine the correct classification. The algorithm for classic AdaBoost for binary problems is shown in Appendix D.1. We will show our algorithm on general binary classification problems which can be easily extended for multiclass problems using the one-against-one technique due to its efficiency. The general algorithm for multiclass problems is shown in Appendix D.2. Extensive experiments on multiclass moving vehicle data show that it is consistently able to select more accurate feature modalities than the classical sequential feature selection method. In order to show the significance of various feature modalities, we'd also like to show a study on individual uni-modal features and their importance for a specified task.

6.3.1 Algorithm for BBFL

First, let us define the notation:

x_i is the ith training sample

y_i is class label of the ith training sample

M is the total number of training samples

$h_t^j(x_i)$ is the label predicted by the *tth* weak classifier $h_t(.)$ for the datum x_i using jth feature subset

w_t is the weight distribution of samples at the *tth* weak learner.

J' is total number of uni-modal features

F^J is the set of all possible uni-modal features and their linear combinations, where J is the total number of feature subsets.

T is the number of weak learners

r_t^j is the overall error rate for the *t-th* weak classifier using *j-th* feature subset

α_t^j is the important factor the *t-th* weak classifier using *j-th* feature subset

$H(x)$ is the final ensemble classifier.

Let $S = (x_i, y_i)_{i=1}^M$ be the set of M training data, s.t. $x_i \in R^D$ and $y_i = \{-1, +1\}$ is the corresponding class label. Let h(.) be a weak classifier which projects an input vector x into {-1, +1} considering only binary classifiers, so that $h_t^j(x_i)$ is the label predicted by the *t-th* weak classifier $h_t(.)$ for the datum x_i using j-th feature subset. This can be applied to any real-valued weak classifiers. In fact, one-against-one multiclass problem is considered as a combination of all binary classifiers. Given J' uni-modal features, F^J is the set of all possible uni-modal features and their linear combinations, where J is the total number of feature subsets. The same weak classifiers *h(.)* are trained on all possible linear combinations of multimodal features at each learning step. A valid weak classifier should have an overall training error rate *r* larger than 0 (meaning more than 50% correctness),

therefore the total number (*J*) of useful feature subset may be different at each iteration. In each step, only the best classifier who has the largest import factor is selected to boost the ensemble classifier. As a result, the best feature set is obtained.

The algorithm for exhaustive BBFL can be described as the follows.

Input: S, T, F^J

Initialize:

t=0; $w_t^i = 1/M$

For t=1 to T:

(1) For all $f^j \epsilon F^J, j = 1, \dots, J$

a. Train a weak classifier $h_t^j(.)$

b. Compute: $r_t^j = \sum_i w_t^i y_i h_t^j(x_i) \ / \ \sum_i w_t^i$

c. Compute: $\alpha_t^j = \log \frac{1-r_t^j}{r_t^j}$

(2) Select the best weak classifier h_t^{j*} who has the largest α_t^{j*}

(3) Re-weight samples: $w_{t+1}^i = w_t^i \exp\left(-\alpha_t^{j*} y_i h_t^{j*}(x_i)\right) / Z_t$, where Z_t is the normalization factor so that $\sum_i w_{t+1}^i = 1$

Output:

An ensemble classifier using the best feature subset

$$H^{j*}(x) = argmax_{j*} \sum_{t=1}^{T} \alpha_t^{j*} h_t^{j*}(x_i)$$

Note that this algorithm is very similar to the classic Adaboost but with an additional feature modality selection at each learning step using the same weak classifier. The weak learning for feature modalities or their combination can be learned independently, and

then the one with best classification accuracy is selected at each step. A MBFS technique can also be employed for a more systematic feature modality selection for every weak learner. However, no matter what weak classifiers or learning techniques used at each learning step, the same number of samples has to be re-weighted and re-evaluated again. Therefore, the BBFL can provide a robust feature modality selection but with increase of the time complexity.

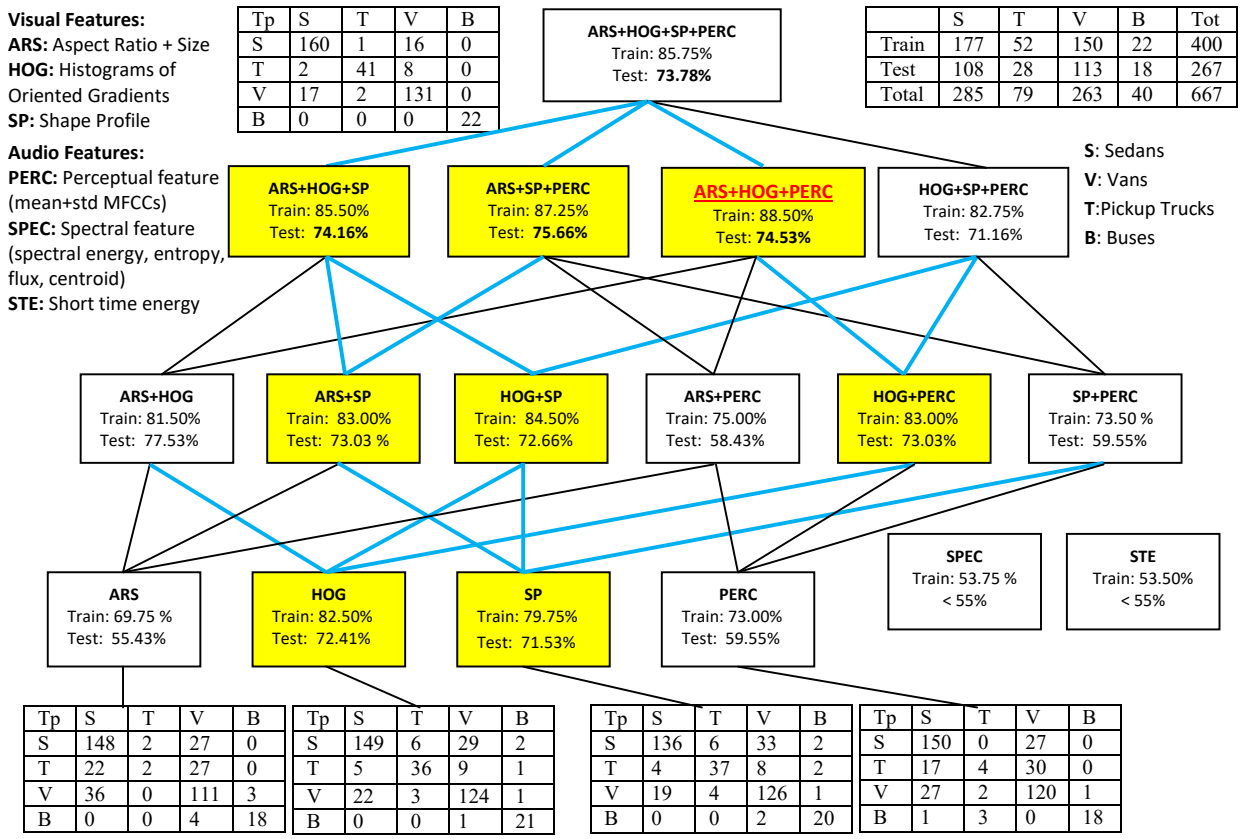
6.4 Experimental Results

We will show our experimental results in two parts. One set of results uses the MBFS technique and another one uses the BBFL technique. We use the same dataset in Chapter 4 of 667 vehicles samples, 400 are used for training and 267 for testing. All vehicles' visual images are reconstructed so that the image results are invariant to perspective views, and the occlusions and motion blurs are removed. The vehicles are labeled into four categories: sedans, vans, pickup trucks and buses. There are various variations in each category. For example, sedans contain sport cars and economic 2 door or 4 door cars (some with fastback or hatchback); vans include mini vans, regular size vans and long size vans (note that SUVs are also categorized into vans as well); pickup trucks some may have wagons or trailers at rear parts; and buses include both school buses and transportation buses.

6.4.1 Results Using MBFS

Figure 6.1 shows the classification results of all the individuals and combinations of multimodal features, including single modalities, bi-modalities, and multimodalities (>=3).

The four vehicle types are labeled as: S-sedan, V-van, T-truck, B-bus. In the training confusion matrices, the ground truth labels are on the rows, and expected labels on the columns. The yellow shading boxes indicate the "good" features that are selected at each level of combinations, and the bold blue lines show their derived branches. Confusion matrices of four meaning single-modal features and the best multimodal features on the same training data are presented in the figure.



Tp	S	T	V	B
S	160	1	16	0
T	2	41	8	0
V	17	2	131	0
B	0	0	0	22

	S	T	V	B	Tot
Train	177	52	150	22	400
Test	108	28	113	18	267
Total	285	79	263	40	667

Tp	S	T	V	B
S	148	2	27	0
T	22	2	27	0
V	36	0	111	3
B	0	0	4	18

Tp	S	T	V	B
S	149	6	29	2
T	5	36	9	1
V	22	3	124	1
B	0	0	1	21

Tp	S	T	V	B
S	136	6	33	2
T	4	37	8	2
V	19	4	126	1
B	0	0	2	20

Tp	S	T	V	B
S	150	0	27	0
T	17	4	30	0
V	27	2	120	1
B	1	3	0	18

Figure 6.1 Comparison of classification results using multimodal features (ARS, HOG, SP, PERC, SPEC, STE and their combinations).

At the bottom level, none of the unimodal features achieve a testing accuracy of over 75%. However, we can observe specific strengths of different features by looking into

their confusion matrices. The ARS feature is the simplest visual feature, therefore it obtains the worst individual performance, but it can help to distinguish vehicles with different sizes and aspect ratios, for example sedans from trucks and buses. Since HOG feature counts the interior structure of vehicles (e.g., windows), its overall performance is the best, but for individual class labels it is not the best (e.g. SP outperforms it in Truck, and PERC in Sedan). The shape profile features analyze the global shape of vehicles and are therefore seems to be the best at distinguishing trucks and other types (particularly vans, since their top rear parts are usually quite different). The PERC feature individually has slightly better performance than the simple ARS feature; from their confusion matrices we can see that it does much better in separating vans from sedans, probably because their sound is more distinct than their aspect ratios and sizes. We will further see how this will make a difference in multimodal integration. The SPEC and STE are not good in combining with other modalities since their training accuracies are less than 55%. Because each modality has its own advantages and disadvantages, the combination among them becomes important to provide complementary information, which we will see next.

In bi-modal classification, we experimented on both visual only and visual and audio cases. In visual feature combinations, HOG and SP are applied on size-normalized images, but their combination includes both interior and exterior information of vehicles, thus providing some classification improvement. ARS feature preserves the size information of vehicle, and therefore providing complementary information to HOG or SP; when combined with either HOG or SP, we also see improvement in testing accuracies,

particularly in ARS+SP. The PERC feature adds acoustic signatures of vehicles in addition to their visual information, thus providing significant improvement over the audio–only results. The testing accuracy using PERC with HOG is slightly better than using HOG itself, indicating features from two different sources (audio and visual) are better than the single source, even though individually, visuals do better than audio.

In the multimodal level, combining 3 or more than 3 features improve the classification. For example, the combination of ARS, HOG and SP (all visual features) increases the accuracy since each of them inherits distinct signature of vehicles. When combining visual features with audio features, the results are also improved. Based on the results, the accuracies with three modalities, between two different visual-audio combinations (ARS+HOG+PERC and ARS+SP+PERC) are very close; the former is slightly better in training and the latter in testing However, SP feature has only 30 dimensions whereas HOG uses 162 dimensions. Therefore, if the reconstructed images are accurate, the SP can be used to replace HOG while combining with other features to reduce computational costs for the vehicle classification task. In fact, the total feature size of ARS+SP+PERC is 63, which is even smaller than the size of the HOG feature vector (162). Between the visual-audio combinations (ARS+HOG+PERC and ARS+SP+PERC) and the visual-only combinations (ARS+HOG+SP), heterogeneous multimodal combinations seem to win, by 3% with the testing set used in this experiment.

At the very top in Figure 6.1, the combination of all useful features has testing accuracy (73.78%), which may not be the best in performance. Therefore, in selecting best

combination of feature modalities, only the one with highest training accuracy is finally chosen. In this experiment, ARS+HOG+PERC is selected with training accuracy 88.50%. Note that its testing accuracy is the second best of the available options.

Here we show an example on how the best multimodal combination wins with the training data. Looking into the training confusion matrices in Figure 8, we have found that buses are misclassified for every one of the four features: ARS (4 misclassification), HOG (1 misclassification), SP (2 misclassification), and PERC (4 misclassification). However, the combination of these four features has 0 misclassifications. This is possible that each feature type has wrong classification on different samples, but by combining all or part of the feature types, those wrong classification results could be corrected. For example, ARS may misclassify samples 2, 3, 4, HOG may misclassify samples 7, 8, 9, and PERC may misclassify samples 10, 11, 12. When combined together there could be no misclassification.

Because the classes the task we used is designed based on visual appearance, the visual feature gave much better results than audio features. Nevertheless, adding audio features to visual feature would provide some improvement, say HOG+SPEC had 2.0% and 0.25% improvement than HOG (the best single modality) itself on both the training set and the testing set, respectively. The best combination, ASR+HOG+PERC, outperformed the best single modality HOG by 6% and 2.12% on the training set and the testing set, respectively.

6.4.2 Results on the Best Feature Combination (ARS+HOG+PERC)

Table 6.1 Training and testing accuracies of ARS+HOG+PERC

Training: 88.50%				
	S	T	V	B
S	160	1	16	0
T	2	41	8	0
V	17	2	131	0
B	0	0	0	22

Testing: 74.53%				
	S	T	V	B
S	85	5	17	1
T	2	24	2	0
V	23	9	80	1
B	4	1	3	10

Based on our experiment results (Figure 6.1), we select the best combination of feature modalities ARS+HOG+PERC based on the training accuracy. Its training and testing confusion matrices and accuracies are show in Table 6.1. Because we use one-against-all for the multi-class classification, we will show the receiver operating characteristic (ROC) curves on each type of vehicles separately in Figure 6.2. In a ROC curve, the true positive rate (sensitivity) is plotted against the false positive rate (1-specificity) for different cut-off points. The sensitivity is the probability that a test result will be positive when the corresponding vehicle type is present; whereas the specificity is the probability that a test result will be negative when the corresponding vehicle type is not present. So, each cut-off point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold which is the estimated probability of a sample calculated from the SVM cost function in order to make a possible decision. A test with perfect discrimination (no overlap in the two distributions) has a ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore, the closer the ROC

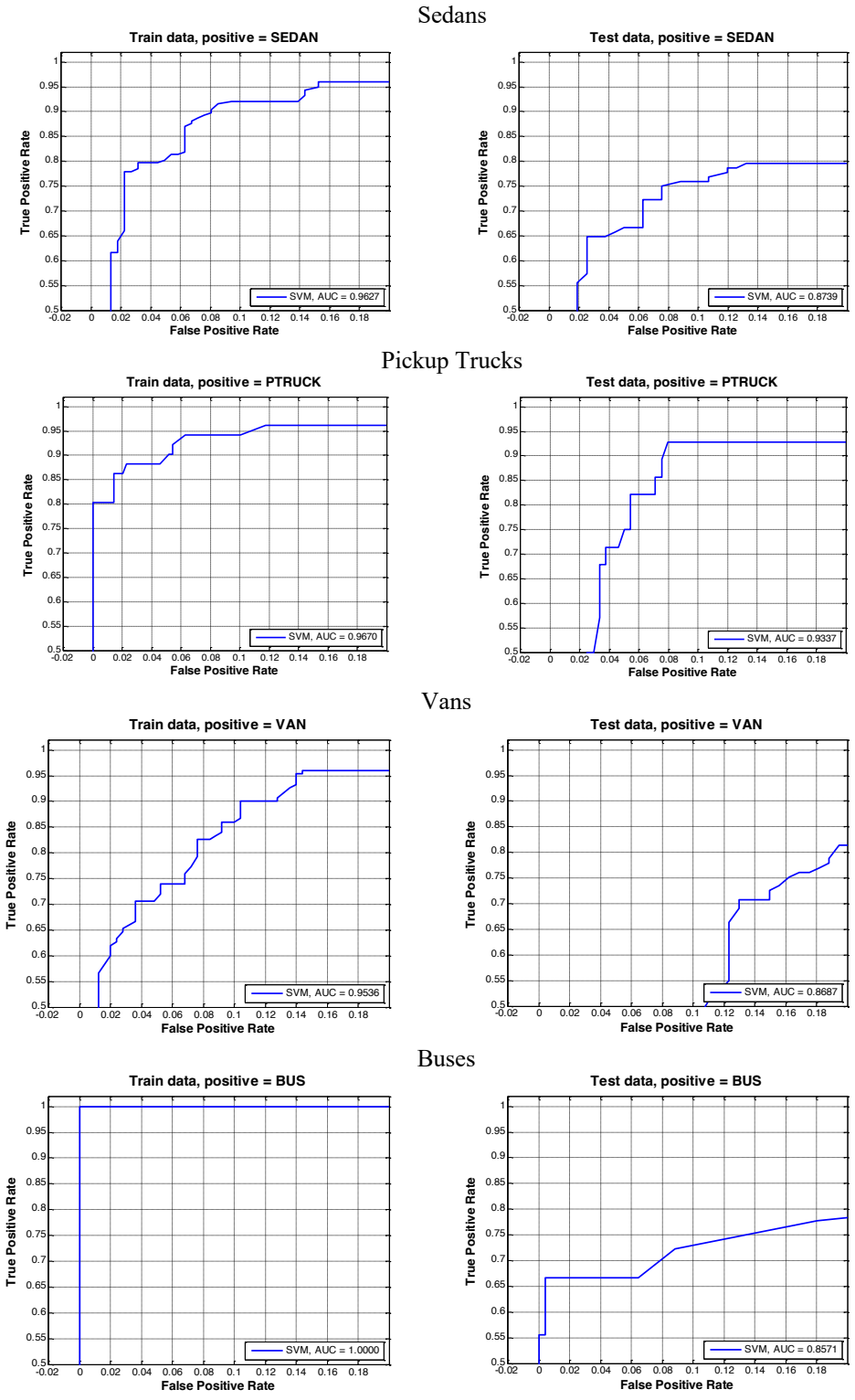


Figure 6.2 ROC curves. All zoomed in on top left corner in the same scale.

curve is to the upper left corner, the higher the overall accuracy of the test. The area

under the ROC curve (AUC) is used to measure the training and testing accuracies of one type against the rest. The Bus class has AUC 1.0 (100%) and all 22 buses are classified correctly in the training, meaning this feature combination can significantly distinguish buses with other types. However, the testing AUC is the lowest for the busses, comparing to the others. That may because the number of samples for buses is much smaller compare to the others. The AUCs for classes with large samples such as sedans and vans also indicated the classifiers are good to separate those types against the rest.

6.4.3 Results Using BBFL

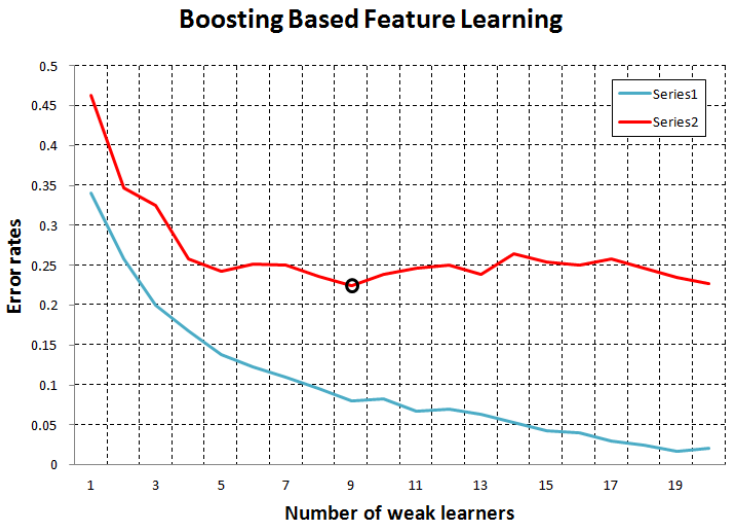


Figure 6.3 Training and testing errors up to 20 weak learners

In the boosting framework, we use decision trees as the weak classifiers. We applied up to 50 weak learners. Figure 6.3 shows the training and testing errors over the number of weak learners. Only the first 20 weak learners are show here. Table 6.2 shows the feature modalities selected at each learning step related to Figure 6.3. The 2nd column shows the accumulated training errors from previous classifiers. The 3rd column shows the testing

errors. The advantages of boosting is that it can select multiple feature modality sets and continuously learn a new feature modality set without stopping at a local maximal one. The last three columns show the 3 best selected of feature modality sets. Note that the training accuracy keeps improving as the number of weak learners increases. However, the best testing error in this experiment (in Figure 6.3) stays at 0.2247. This is because the decision nodes selected depend only on the re-weight samples in training. The accuracy of those testing samples only depends on the number of possible feature modalities selected. For example, in this experiment in Table 6.2, the first three weak learners select the same feature combination: ARS+HOG. Since they only use visual signatures, the testing errors are the worst. No. 4 to No. 6 weak learners select the same feature combination: ARS+HOG+PERC. They include some acoustic signatures, so the testing errors decrease some. Then No. 8 and No. 9 weak learners select the combination of the most representative uni-modal feature modalities: ARS, HOG, PERC and SPEC, and those could provide the most complementary information to each other. And using No. 9 weak learners gives the best testing accuracy among all 20 weak learners (in Figure 6.3). Note that when different feature combinations are selected later, such as using No. 12 to No. 18 weak learners, the testing errors actually increase. Until No. 19 and No, 20 weak learners are used, the testing errors decrease again closing to the error using No. 9 weak learner. The training and testing errors as well as the confusion matrices using No. 9 weak learners are show in **Error! Reference source not found.**. Note that the number of feature modalities selected at the 2nd best (as well as the 3rd best) is larger than the top best in training at most weak learning steps. Therefore, it is important to select feature

combinations with less number of modalities in order to reduce the amount time in testing.

Table 6.2 The first 20 iterations of boosting-based feature modality learning

T	Train	Test	Top Best	2nd Best	3rd Best
1	0.3409	0.4627	ARS+HOG	ARS+HOG+STE	ARS+HOG+PERC
2	0.2581	0.3470	ARS+HOG	ARS+HOG+STE	ARS+HOG+SPEC
3	0.2005	0.3246	ARS+HOG	ARS+HOG+STE	ARS+HOG+PERC
4	0.1679	0.2575	ARS+HOG+PERC	ARS+HOG+STE+PERC	ARS+HOG+PERC+SPE
5	0.1378	0.2425	ARS+HOG+PERC	ARS+HOG+STE+PERC	ARS+HOG+PERC+SPE
6	0.1228	0.2512	ARS+HOG+PERC	ARS+HOG+STE+PERC	ARS+HOG+PERC+SPE
7	0.1103	0.2497	ARS+HOG+SPEC	ARS+HOG+STE+SEPC	ARS+HOG+PERC+SPE
8	0.0952	0.2359	ARS+HOG+PERC +SPEC	ARS+HOG+STE+PERC +SPEC	ARS+HOG+PERC
9	0.0800	0.2247	ARS+HOG+PERC +SPEC	ARS+HOG+STE+PERC +SPEC	> Err
10	0.0827	0.2388	ARS+HOG+PERC +SPEC	ARS+HOG+STE+PERC +SPEC	> Err
11	0.0677	0.2463	ARS+HOG+PERC +SPEC	ARS+HOG+STE+PERC +SPEC	> Err
12	0.0702	0.2500	ARS+HOG+SPEC	ARS+HOG+STE+SEPC	ARS+HOG+PERC
13	0.0627	0.2388	ARS+HOG+PERC	ARS+HOG+STE+PERC	> Err
14	0.0526	0.2649	ARS+HOG+SPEC	ARS+HOG+STE+SEPC	> Err
15	0.0426	0.2537	ARS+HOG+PERC	ARS+HOG+SPEC	ARS+HOG+STE+PERC
16	0.0401	0.2500	ARS+HOG+STE +SEPC	ARS+HOG+SPEC	> Err
17	0.0301	0.2575	HOG+PERC+SPEC	ARS+HOG+PERC+SPEC	ARS+HOG+STE+PERC +SPEC
18	0.0251	0.2463	ARS+HOG+SPEC	ARS+HOG+STE+SEPC	> Err
19	0.0175	0.2351	ARS+HOG+PERC +SPEC	ARS+HOG+STE+PERC +SPEC	> Err
20	0.0201	0.2276	ARS+HOG+PERC +SPEC	ARS+HOG+STE+PERC +SPEC	> Err

Table 6.3 The best testing results of the boosting based feature learning using 9 weak learners.

Training: 92.00%				
	S	T	V	B
S	165	1	11	0
T	3	45	3	0
V	11	3	136	0
B	0	0	0	22

Testing: 77.53%				
	S	T	V	B
S	86	1	21	0
T	2	20	6	0
V	16	5	91	1
B	3	0	5	10

6.4.4 Comparison Between MBFS and BBFL

The main difference between the MBFS and BBFL algorithms is that the MBFS only selects one best combination of feature modalities whereas BBFL selects many feature modality sets. The MBFS starts with selecting the best feature modality from all uni-modal features then combine it with those not selected in the next step. These procedures are repeated until combinations of all feature modalities are evaluated. The worst computation is when all feature modalities or their combinations have similar classification accuracies and fall into the decision boundary, so that all are selected at each level (or each round). So the time complexity for MBFS is $C_{SVM}\sum_{k=1}^{n}\binom{n}{k}$, where $\sum_{k=1}^{n}\binom{n}{k}$ is the total number of all possible combinations of *n* uni-modal feature modalities, and C_{SVM} is the time to evaluate a feature modality using the SVM. This assumes that all feature modalities have same number of vector dimensions. This is always not true. For example, ARS has only 3 dimensions, whereas HOG has 162 dimensions in the previous results, so the time to train using ARS feature is much faster than that using HOG feature for the SVM. The BBFL evaluates the same number of re-weighted samples using a number of weak learners. So if the MBFS technique is employed at each weak learning step, the time complexity for the BBFL will be $TC_{DT}\sum_{k=1}^{n}\binom{n}{k}$, where T is the number of weak learners and C_{DT} is the time to evaluate a feature modality using the decision tree weak classifier. Even though the classification time using decision trees is much faster than using SVM on individual feature modalities, the large number (T) of the weak learners will make the computation expensive using the BBFL in feature modality selection. In our experiments as shown

above, the time to select the best combination of feature modalities using the MBFS is about 0.61 seconds, whereas the time to learn all feature modality sets of 50 weak learners using the BBFL is about 4.76 seconds, which is 7 to 8 times slower. The computer that we used has Intel CPU 3.06GHz with installed 4GB memory. So, if the training criteria are not met, larger number of weak learners could be used in order to obtain robust results, then the computational time increases.

For the classification performance, the selected best classifier with the MBFS technique achieves a training accuracy of 88.50% and a testing accuracy of 74.53%. The selected feature combination is ARS+HOG+PERC. The best performance with the BBFL technique achieves a training accuracy of 92.00% and a testing accuracy of 77.53%. The testing accuracy is 3% higher than that of the MBFS, but this is achieved with the ensemble of 9 weak learners. Notably, among the 9 weak learners, the most important modalities are ARS+HOG+PERC, which is consistent with the results using the MBFS technique.

6.5 Concluding Remarks

In this chapter, various multimodal features are systematically integrated and studied for vehicle classification. Results show that using multimodal features can have significant improvement in classification performance over that using single modality. We also make a number of important observations on the strengths and weakness of various features and their combinations. In addition, for some types of features the combination is computationally faster than a complicated feature, with similar classification performance. Two techniques are proposed for feature modality selection. The MBFS

selects the best combination of feature modalities, whereas the BBFL selects many combinations. The BBFL is more robust in using ensemble of many weak learners, however it is computational expensive than the MBFS. In the end, we would like to point out that those algorithms are based on our reconstructed visual data and filtered audio data. In other words, high quality detection often over-weights the choice of algorithms, since more distinctive and stable features can be selected. So, if we have a large high-quality dataset, the choice of classification algorithms might not really matter so much in terms of classification performance. Therefore, feature modality selection becomes very important in terms of selecting sensor sources and data to be collected.

Chapter 7

7 Conclusions and Future Work

This thesis presents a framework for multimodal sensing and process for moving object detection and classification. We used the dataset contains moving vehicles as the particular example throughout the thesis. The proposed approaches build upon novel ideas in sensor designs, image and video processing, signal processing and machine learning to provide general methods for feature modality selection and object classification. The main contribution of our work is the unified Adaptive and Integrated Multimodal Sensing and Processing (AIM-SP) framework to integrate sensing, feature selection and classification. A number of papers related to this thesis have been published in journals and conferences (Qu, et al., 2010; Li, et al., 2010; Wang., et al, 2010b; Wang., et al., 2011a; Wang, et al., 2011b; Wang, et al., 2012a; Wang, et al., 2012b), and more are under review and preparation; for a complete list, please see Appendix E.

We will summarize our key contributions in Section 7.1, discuss some limitations in Section7.2 and propose some future research directions in Section 7.3.

7.1 Key Contributions

Within this framework, three unique contributions are made:

A novel Vision-Aided Automated Vibrometry (VAAV) multimodal sensor system. This system is built upon a novel sensor technology, LDV, and is capable of obtaining

visual, range and acoustic signatures for moving object detection at a large distance. The system consists of a dual-PTZ camera based stereo vision system for improving the automation and time efficiency of LDV long-range remote hearing. The closed-loop adaptive sensing using the multimodal platform allows determination of good surface points and quickly focusing the laser beam based on the target dtection, surface point selection, distance measurements, and LDV signal returning feedbacks. The integrated system greatly increases the performance of the LDV remote hearing and therefore its feasibility for audio-visual surveillance and long-range other inspection and detection applications.

A multimodal temporal panorama (MTP) approach for moving object detection and extraction. The MTP integrates visual appearance, motion information and acoustic signals of moving vehicles for multimodal data representation and alignment. The technique of using a vertical detection line and a horizontal epipolar line can detect a moving vehicle efficiently in real time. The MTP also helps in data labeling effectively, especially with a large amount of data. In addition, it provides the capability to reconstruct vehicles' visual appearances so that motion blurs, occlusions and perspective distortions can be removed. It also provides a very effective user interface for training data labeling in both video and audio domains. The concept of MTP is not limited to visual and audio information, but is also applicable when other modalities are available that can be presented in same time axis.

Feature modality selection using a multi-branch feature searching (MBFS) technique and a boosting based feature learning (BBFL) technique. Multimodal features can have significant improvement in classification over that using single modality. The MBFS selects the best combination of feature modalities with high time efficiency, whereas the BBFL selects many combinations with high robustness. Base on the experimental results, a number of important observations on the strengths and weakness of various features and their combinations are made as well.

In addition, a new audio visual vehicle (AVV) dataset is created for moving object detection, classification, and potentially identification.

7.2 Limitations of Our Approaches

Our multimodal sensing system targets on moving object detection and classification. It assumes the target can be detected from at least one modality, either visual, audio or motion. However, the current fusion system assumes that all modalities are available, so it remains a future research if one of the modalities is missing. For example, an electric car which does not make engine sound can only be detected if it can be observed in video.

The moving object also needs to have a speed to be extractable using the MTP approach, for example, vehicles or people riding bikes. It can also detect a pedestrian, but it is not good at reconstructing it. Figure 7.1 shows examples of two walking people are reconstructed based on our MTP technique. Because the walking speed of a person is very slow, the number of pixels extracted from the original image shots is limited. The

integration of multiple slices will cause some body parts missing in the reconstructed results.

Figure 7.1 Pedestrian reconstruction

If multiple moving objects overlap with each other, the detection and reconstruction also have problems. For example, in Figure 7.2, two people were riding bikes while being overlapped by a moving vehicle. If the reconstruction is based on the speed of those people, the vehicle shape cannot be fully recovered. So, determining which speeds of any two objects to be used from the motion slopes that are mixed together in the EPI is a tricky problem.

Figure 7.2 Reconstruction results of people on bikes and a moving

Assume all objects are correctly reconstructed and features can be extracted. Then, feature modality selection provides an advantage to select the most representative

features for a specific task. It also can give a feedback of what kind of sensors are important to continue collecting data for the task. However, when multiple sensors are employed and many features are extracted, the training will take time to learn all possible feature combinations. Especially in BBFL, all possible feature combinations need to learn again for every time a new weak learner is used. The time complexity is dramatically increased when a large number of weak learners are used.

7.3 Future Work

This section provides some discussion on future research in multimodal sensing and processing for moving object detection and recognition.

To solve the illumination problem of the PTZ cameras, an IR sensor can be included in the sensing system to capture the data at a poor lighting condition. This provides additional object signatures that can improve the object detection. However, when to use the IR sensor for object detection should be carefully considered. The additional sensor modality can increase the latency in data capturing and synchronization. More feature modalities can be extracted to improve the object detection and classification but the time complexity in feature learning and selection also increases.

There is always a case when not all feature modalities are available during either training or testing. The cross modal learning can provide the ability to learn using the available set of feature modalities and test on the other types. However, in the modeling of the cross-modal learning, representation and integration are two key problems. So, which

modalities or sub-modalities are to be represented, and how they can be integrated are the most interest.

The data we collected can also be used for vehicle recognition and anomaly detection. Some same vehicles were appeared at different time and days during data collection. It would be interested to monitor the behaviors of vehicles and recognize the same one over days. For example, Figure 7.3 shows possibly a same moving vehicle passed through the monitoring area at different time and days. The original images (at top row) show the vehicles moving at different directions and locations. It is hard to match them and identify them as the same vehicle. However, after reconstruction, their shapes, even the people inside the vehicles, are very similar that can indicate a possibility of the same one.

Figure 7.3 Same vehicle at different time and days by comparing the reconstruction results.

The multimodal sensing framework is not limited to moving vehicle detection only, but also can be applied for general applications, such as surveillance, check point inspection, searching and rescue. The general idea behind it is: given multiple sensor sources we want

to select the most important sensor modalities from multimodal data analysis and feature modality learning. More criteria should be studied and evaluated in the future work.

Appendix A: PTZ and LDV Calibration

Assume the intrinsic parameters of the PTZ are already calibrated and known from the calibration of the two PTZ cameras. Then we only focus on solving the extrinsic parameters between the PTZ and the LDV.

According to equation 3.4 and 3.7, we have

$$\begin{cases} P_L = R_{C'} P_{C'} + T_{C'} \\ P_L = R_U R_{LR} R_U^T (P_{ML} - T_U) + T_U \end{cases} \quad (A.1)$$

Then,

$$R_{C'} P_{C'} = R_U R_{LR} R_U^T (P_{ML} - T_U) + (T_U - T_{C'}) \quad (A.2)$$

Note that R_U contains three rotations matrices: an initial rotation matrix R_{U0}, a pan rotation matrix $R_{U\alpha}$, and a tilt rotation matrix $R_{U\beta}$. Assume the initial rotation R_{U0} is unit matrix, we have

$$R_{C'} P_{C'} = R_k (P_{ML} - T_U) + (T_U - T_{C'}) \quad (A.3)$$

where $R_k = R_{U\alpha} R_{U\beta} R_{U\beta}^T R_{U\alpha}^T$

Let,

$$R_{C'} = \begin{pmatrix} rc_{11} & rc_{12} & rc_{13} \\ rc_{21} & rc_{22} & rc_{23} \\ rc_{31} & rc_{32} & rc_{33} \end{pmatrix}, \; R_k = \begin{pmatrix} rk_{11} & rk_{12} & rk_{13} \\ rk_{21} & rk_{22} & rk_{23} \\ rk_{31} & rk_{32} & rk_{33} \end{pmatrix} \quad (A.4)$$

Note that in the mirrored coordinate system P_{ML}, a point P always has the form of $[0,0,t_l]$ since it is on the Z axis. Thus, Eq. A.3 can be expanded using A.4 as:

$$\begin{pmatrix} rc_{11} & rc_{12} & rc_{13} \\ rc_{21} & rc_{22} & rc_{23} \\ rc_{31} & rc_{32} & rc_{33} \end{pmatrix} \begin{pmatrix} X_C \\ Y_C \\ Z_C \end{pmatrix} - \begin{pmatrix} rk_{11} & rk_{12} & rk_{13} \\ rk_{21} & rk_{22} & rk_{23} \\ rk_{31} & rk_{32} & rk_{33} \end{pmatrix} \begin{pmatrix} 0 - T_{Ux} \\ 0 - T_{Uy} \\ t - T_{Uz} \end{pmatrix} - \begin{pmatrix} T_{Ux} - T_{Cx} \\ T_{Uy} - T_{Cy} \\ T_{Uz} - T_{Cz} \end{pmatrix} = 0$$

(A.5)

In Eq. A.5, the variables $t - T_{Uz}$and $T_{Uz} - T_{Cz}$ are dependent. Thus, the equation is non-linear. In order to solve this problem, assume $T_{Uz} = d$ as a constant, the variables $t - T_{Uz}$ and $T_{Uz} - T_{Cz}$ will become independent, then the equation turns into linear. Consequently, Eq. A.5 can be written a homogenous equation that includes n+14 unknowns. Each calibrating point can built 3 equations. So we need at least 7 unique points to solve the Eq. A.5. However, this approach is very sensitive to noise of the captured data. Therefore, the platform needs to be calibrated using multiple steps. First, make the initial matrix in R_C a unit matrix, both T_{Cy} and T_{Uy} equal to 0 thus solves T_{Cx}, T_{Cz} and T_{Ux}. Second, still make the initial matrix in R_C a unit matrix, solve T_{Cy} and T_{Uy} after obtain T_{Cx}, T_{Cz} and T_{Ux}. Third, refine the initial matrix in R_C after obtain the translation vectors T_C and T_U. Finally, refine T_U according to the calibration relation between distances and focus steps of the LDV.

Appendix B: Laser Camera Alignment

Recall the geometric modal in Figure 3.8, $\overrightarrow{\boldsymbol{BA}}$ is the laser beam from the LDV, and $\overrightarrow{\boldsymbol{AD}}$ is the reflected laser ray from the mirror which has its normal along AC after pan α and tilt β. AE is the project of the AD on the plane ABE. We need to solve $\angle EAF = \alpha'$ and $\angle DAE = \beta'$.

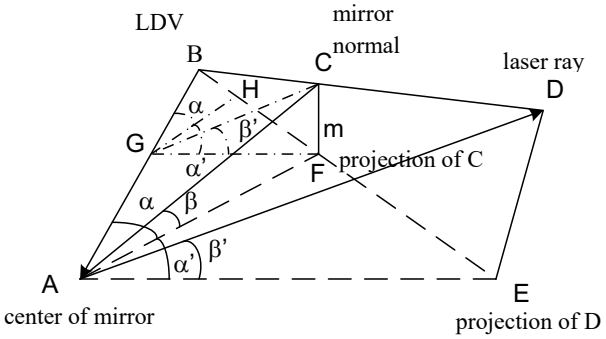


Figure 3.8. Geometric model of laser beam from the LDV ($\overrightarrow{\boldsymbol{BA}}$) and its reflected laser ray ($\overrightarrow{\boldsymbol{AD}}$) after the pan ($\alpha$) and tilt ($\beta$).

Here is the detailed derivation;

Step 1. Draw GC || AD, GH $\perp$ BE, since GF is the project of GC and AE is the project of AD, thus, GF || AE. Let CF = m. The problem is turned to solve $\angle FGH = \alpha'$ and $\angle CGF = \beta'$.

Step 2. $\tan(\alpha') = \tan(\angle FGH) = FH / GH$

$\sin(\beta') = \sin(\angle CGF) = CF / GC = m / GC$

We need to solve GC, FH and GH

Step 3. $\mathrm{GC} = \mathrm{AG} = \frac{1}{2} * \frac{\mathrm{AC}}{\cos\alpha * \cos\beta} = \frac{1}{2} * \frac{\mathrm{m}}{\sin\beta} * \frac{1}{\cos\alpha * \cos\beta}$

$$= \frac{\mathrm{m}}{2 * \sin\beta * \cos\alpha * \cos\beta} = \frac{\mathrm{m}}{\sin 2\beta * \cos\alpha}$$

Step4.

$$\mathrm{FH} = \mathrm{BF} - \mathrm{BH} = \mathrm{AF} * \tan\alpha - \mathrm{BG} * \sin\alpha$$

$$= \mathrm{m} * \tan\alpha * \cot\beta - (\mathrm{AB} - \mathrm{AG}) * \sin\alpha$$

$$= \mathrm{m} * \tan\alpha * \cot\beta - \left(\frac{\mathrm{m}}{\cos\alpha * \tan\beta} - \frac{\mathrm{m}}{\cos\alpha * \sin 2\beta}\right) * \sin\alpha$$

Step 5.

$$\mathrm{GH} = \mathrm{BG} * \cos\alpha = (\mathrm{AB} - \mathrm{AG}) * \cos\alpha$$

$$= \left(\frac{\mathrm{m}}{\cos\alpha * \tan\beta} - \frac{\mathrm{m}}{\cos\alpha * \sin 2\beta}\right) * \cos\alpha$$

As a result,

$$\alpha' = \tan^{-1}\left(\frac{\tan\alpha}{\cos 2\beta}\right)$$

$$\beta' = \sin^{-1}(\cos\alpha * \sin 2\beta)$$

Appendix C: Reconstructed Image Results

Here we show more reconstructed image results selected from the dataset of 667 sample vehicles. The image resolution captured from the PTZ is 720x480. The data are capture at two locations, one at a local road about 25-30 meters, one at a highway about 50-70 meters. The data collected at local road use two camera zoom levels, simply speaking, zoom in and zoom out views.

Figure C.1 shows the 2 door sports car collect using a zoom out view. The reconstructed image size of the first one is 182x59 in pixels. The reconstructed image size of the second one is 180x59 in pixels

Figure C.1 2-Door sports cars. Original image shots on top, reconstruction resutls on bottom

Figure C.2 shows some special 4 door sedans. The first row shows a regular black 4 door sedan. The reconstructed image size is 180x57. The second row shows a special 4 door sedan with a station wagon at rear part. The reconstructed image size is 221x61. The third row shows the car with long board in the truck. The reconstructed image size for that one is 210x53. The last row shows the car with a cart at tail. The reconstructed image size for

that one is 338x65. Therefore, the variations of different kinds of sedans may cause the misclassification into other types of vehicles, say, vans or pickup trucks, etc.

Figure C.2. Different 4-door sedans. Original image shots on left, reconstruction resutls on right

Similar to sedan-type vehicles, some van-type vehicles in Figure C.3 also have a lot of variations. The class vans include mini-vans, SUVs, jeeps, regular vans, and long vans. Other types of vehicles such as pickup truck and buses are shown in Figure C.4.

Figure C.3. Different van-type vehicles. Original image shots on left, reconstruction resutls on right

Figure C.4. Examples of a pickup and a transportation bus. Original image shots on left, reconstruction resutls on right

Note that our reconstruction technique can also be applied to a dense traffic where many vehicles move closely. Figure C.5 shows two sample scenarios where two vehicles are moving closely. Note that we assume multiple vehicles move at similar speed and same direction in order to reconstruct good results. But it would be still challenging to separate them if no additional information is acquired.

Figure C.5 Date collected at a highway where vehicles move closely. Original image shots on left, reconstruction resutls on right

Appendix D: Boosting Algorithms

D.1 Classic AdaBoost for a Binary Classification Problem

Adaboost was first introduced by Freund and Schapire (1997). In a two-class classification setting, we have training samples $\{\{x_i, y_i\}, i=1,\ldots m\}$ with x_i belongs to a feature domain and $y_i \in \{-1,+1\}$. The procedure of the algorithm is:

1. *Initialize the weight* $w_i=1/m$.
2. *Repeat for* $t=1,2,\ldots,T$:
    a. *Fit the classifier* $h(x_i) \in \{-1,+1\}$ *using weights* w_i *on the training data.*
    b. *Compute the* $\varepsilon_t = Prob(h_t(x_i) \neq y_i)$, $\alpha_t = ln(1/\varepsilon_t - 1)$.
    c. *Update the weights:*

$$w_i \leftarrow w_i \exp(-\alpha_t y_i h_t(x_i))/z_t$$

where z_t *is normalization factor.*

3. *Output the final classifier:*

$$H(x) = \Sigma_t \alpha_t h_t(x)$$

D.2 Boost for a K-Class Classification Problem

Given a training samples $\{\{x_i, y_i\}, i=1,\ldots m\}$, where x_i belongs to a feature domain and y_i is the label of instance i. Let $U^k=\{u_1,\ldots u_k\}$ be the unit base vectors of a K-dimension space R^K. The labels can be represented by one of the base vectors, i.e., $y_i=u_n$ if the instance i is the class n. The procedure of the algorithm is:

1. *Initialize the weight w_{ij} (i=1,…,m and j=1,…,K):*

$$w_{ij} = \begin{cases} 0 & y_i = u_j \\ 1 & otherwise \end{cases}$$

2. *For t=1,…,T:*

    a. *Normalize w_{ij}.*

    b. *Train $h_t(x)$ by minimizing loss function:*

    $$L = \sum_{i=1}^{m} \sum_{j}^{k} w_{ij} exp(\left(u_j - y_i\right) h(x_i))$$

    c. *Update the weight matrix w_{ij}:*

    $$w_{ij} \leftarrow w_{ij} exp(\left(u_j - y_i\right) h(x_i))$$

3. *Final classifier:*

$$H(x) = \sum_{t=1}^{T} h_t(x)$$

Appendix E: Candidate's Publication List

Journals & Book Chapters:

1. **T. Wang**, and Z. Zhu, Vision-Aided Automated Vibrometry for Remote Audio-Visual-Range Sensing, *Smart Sensor Technologies*, *eds. K. Iniewski and M. Syrzycki*, CRC Press, in press. (Invited)

2. R. Li, **T. Wang**, Z. Zhu, and W. Xiao, Vibration Characteristics of Various Surfaces Using an LDV for Long-Range Voice Acquisition*, IEEE Sensor Journal*, vol 11, no 6, June 2011, pp 1415 - 1422.

3. Y. Qu, **T. Wang**, and Z. Zhu, Vision-aided Laser Doppler Vibrometry for Remote Automatic Voice Detection *, IEEE/ASME Transactions on Mechatronics,* issue: 99, pgs. 1-10, November, 2010.

4. **T. Wang**, Z. Zhu, R. S. Krzaczek, and H. E. Rhody, A System Approach to Adaptive Multimodal Sensor Designs, book chapter 7 *of Machine Vision Beyond Visible Spectrum, eds. R. Hammond, G. Fan, R. McMillan, and K. Ikeuchi*, Springer, September, 2010. (Invited)

5. **T. Wang**, Z. Zhu, and E. Blasch, Bio-Inspired Adaptive Hyperspectral Imaging for Target Tracking, *IEEE Sensors Journal, Special issue on Enhancement Algorithms, Methodologies & Technology for Spectral Sensing*, vol. 10, no. 3, March 2010, pp. 647-654.

Conferences:

6. **T. Wang** and Z. Zhu, Multimodal and Multi-task Audio-Visual Vehicle Detection and Classification, 9th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS), Sept. 2012.

7. **T. Wang**, Z. Zhu, A multimodal temporal panorama approach for moving vehicle detection, reconstruction and classification, *conference on SPIE Defense, Security and Sensing*, 8389-30, April, 2012.

8. **T. Wang**, Z. Zhu, Real Time Moving Vehicle Detection and Reconstruction for Improving Classification, *IEEE Computer Society's Workshop on Applications of Computer Vision (WACV)*, 2011.

9. **T. Wang**, Z. Zhu, and C. Taylor, Multimodal Temporal Panorama for Moving Vehicle Detection and Reconstruction, International *Workshop on Video Panorama (IWVP)*, Dec. 2011.

10. **T. Wang**, R. Li, Z. Zhu, and Y. Qu, Active Stereo Vision for Improving Long Range Hearing Using a Laser Doppler Vibrometer , *IEEE Computer Society's Workshop on Applications of Computer Vision (WACV)*, 2011.

11. Y. Qu, **T. Wang** and Z. Zhu, An Active Multimodal Sensing Platform for Remote Voice Detection, *IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM'10)*, 2010.

12. **T. Wang**, Z. Zhu, and A. Divakaran, Long-Rang, Audio and Audio-Visual Event Detection Using a Laser Doppler Vibrometer, *conference on SPIE Defense, Security and Sensing: Evolutionary and Bio-Inspired Computation: Theory and Applications IV*, April, 2010

13. **T. Wang**, Z. Zhu, and H. Rhody, A Smart Sensor with Hyperspectral/Range Fovea and Panoramic Peripheral View, *IEEE CVPR workshop on, Object Tracking and Classification Beyond the Visible Spectrum (OTCBVS)*, Florida, USA, June 20-25, 2009

14. Y. Qu, **T. Wang**, and Z. Zhu, Remote Audio/Video Acquisition for Human Signature Detection, *IEEE CVPR Workshop on Biometrics*, Florida, USA, June 20-25, 2009

15. **T. Wang**, and Z. Zhu, Intelligent Multimodal and Hyperspectral Sensing for Real-Time Moving Target Tracking, *IEEE Applied Imagery Pattern Recognition (AIPR) Workshop 2008,* Washington DC, USA, 2008.

16. **T. Wang**, and Z. Zhu, Bio-Inspired Adaptive Hyperspectral Imaging for Target Tracking, *Symposium on Spectral Sensing Research (ISSSR),* June 23-27, 2008.

Others:

17. **T. Wang**, Z. Zhu, Vision-Aided Automated Vibrometry for Remote Acoustic and Vibration Sensing, CMOS Emerging Technologies, Vancouver, Canada, July, 2012.

18. **T. Wang**, Z. Zhu, Y. Qu, and A. Divakaran, Long-distance audio event detection using a laser Doppler vibrometer, *SPIE Newsroom*, 2010 ,

(http://spie.org/x40847.xml?highlight=x2420&ArticleID=x40847)

Submitted:

T. Wang, Z. Zhu and A. Divakaran, A Survey on Multimodal Surveillance: Sensing, Fusion and Event Recognition, *Computer Vision and Image Understanding (CVIU), Special issue on Advances In Machine Vision Beyond Visible Spectrum,* 2012. (Major Revision)

T. Wang, Z. Zhu and C. N. Taylor, A Multimodal Temporal Panorama Approach for Moving Vehicle Detection, Reconstruction and Classification, *Computer Vision and Image Understanding (CVIU), Special issue on Advances In Machine Vision Beyond Visible Spectrum,* 2012. (Minor Revision)

List of Figures

List of Tables

References

Aleksic, P. S., Williams, J. J., Wu, Z., & Katsaggelos, A. K. (2002). Audio-visual speech recognition using mpeg-4 compliant visual features. *EURASIP J. Appl. Signal Processing*, vol. 2002, no. 11, pp. 1213-1227, November.

Atrey, P. K., Kankanhalli, M. S., & Jain, R. (2006). Information assimilation framework for event detection in multimedia surveillance systems. *Multimedia Systems*, vol. 12, no. 3, pp. 239-253, September

Beal, M. J., Jojic, N., & Attias, H. (2003). A graphical model for audiovisual object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25, 828-836

Bengio, S. (2003). Multimodal authentication using asynchronous HMMs. *In Proc. 4th Int. Conf. Audio- and Video-Based Biometric Person Authentication,* Guildford, U.K., pp. 770-777

Blackmon, F. A. and Antonelli, L. T. (2006). Experimental detection and reception performance for uplink underwater acoustic communication using a remote, in-air, acousto-optic sensor, *IEEE J. Oceanic Engineering*, 31(1), 179-187, January.

Boiman, O. & Irani, M. (2005). Detecting irregularities in images and in video. In *Proc. IEEE International Conference on Computer Vision*, pp. 1985-1988, Beijing, China, Oct. 15-21.

Bolles, R. C., Baker, H. H., & Marimont, D. H. (1987). Epipolar plane image analysis: An approach to determine surface from motion. *Int. J. Computer Vision*, vol. 1, no. 7, 7-15

Bouguet, J. Y. (2008). Camera calibration toolbox for Matlab. Available at: http://www.vision.caltech.edu/bouguetj/calib_doc/index.html, June 2008

Chang, C-C. & Lin, C-J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1—27:27.

Chen, W. , Oetomo, S. B. , Feijis, L., Andriessen, P., Kimman, F., Gerates,M. & Thielen, M. (2010). Rhythm of life aid (ROLA): An integrated sensor system for supporting medical staff during cardiopulmonary resuscitation (CPR) of newborn infant, *IEEE Trans. Information Technology in Biomedicine*, issue 99, May.

Chetty, G. and Wagner, M. (2006) Audio-visual multimodal fusion for biometric person authentication and liveness verification. *In: NICTA-HCSNet Multimodal User Interaction Workshop*, pp. 17-24. Sydney.

Chu, S. M. & Huang, T. S. (2007). Audiovisual speech recognition*. In Z. Zhu and T. S. Huang (Eds.), Multimodal Surveillance: Sensors, Algorithms, and Systems*. pp. 109-140. Norwood, MA: Artech House.

Clavel, C., Ehrette, T., & Richard, G. (2005). Event detection for an audio-based surveillance system. *In Proc. of IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1305-1309

Codec, M., Leistner, C., Bischof, H., Starzacher, A. & Rinner, B. (2010). Audio-visual co-training for vehicle classification, *7th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS).*

Comanicu, D. & Meer, P. (2002). Mean shift: a robust approach toward feature space analysis,". *IEEE Trans. Patten Analysis and Machine Intelligence.* May

Conaire, C. O., O'Connor, N. E., Cooke, E., & Smeaton, A. F. (2006). Multispectral object segmentation and retrieval in surveillance video. *IEEE International Conference on Image Processing*, October

Crebolder, J. M., Unruh, T. D. M., & Mcfadden S. (2003). Search performance using imaging displays with restricted field of view. *Tech. Rep., DRDC Toronto TR 2003-007, Defense R&D* Canada, April.

Cristani, M., Bicego, M. & Murino, V. (2007). Audio-visual event recognition in surveillance video sequences, *IEEE Trans. Multimedia*, 9(2): 257-267, February.

Cui, Y., Samarasekera, S., Huang, Q., & Greiffenhagen, M. (1998). Indoor monitoring via the collaboration between a peripheral sensor and a foveal sensor. *In Proceedings of the IEEE Workshop on Visual Surveillance*, 2-9

Dalal, N. & Triggs, B. (2005). Histograms of oriented gradients for human detection. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dalal, N., Triggs, B. & Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. *In Proceedings of the European Conference on Computer Vision*, Graz, Austria, vol. II, pp. 428-441. May.

Danielsson, D., Rasolzadeh, B, and Carlsson, S. (2011). Gated classifiers: Boosting under high intra-class variation. *In CVPR*, pages 2673-2680, 2. 5

Davis, J. W. & Sharma, V. (2005). Fusion-based background-subtraction using contour saliency. *Computer Vision and Pattern Recognition*, 20-26, June

Dedeoglu, Y., Toreyin, B. U., Gudukbay, U. & Cetin, A. E. (2008). Surveillance using both video and audio, *in Multimodal Processing and Interaction: Audio, Video, Text*, P. Maragos, A. Potamianos and P. Gros Eds., 143-156.

Divakaran, A. (Ed.). (2009). Multimedia content analysis: Theory and applications". *Signals and Communication Technology*, Springer US, March.

Flora, G. & Zheng, J. Y. (2007). Adjusting route panoramas with condensed image slices. *ACM Conf. Multimedia* 07, 815-818, Germany.

Foresti, G., Micheloni, C., Snidaro, L., Remagnino, P., & Ellis, T. (2005). Active video-based surveillance system: the low-level image and video processing techniques needed for implementation. *IEEE Signal Processing Magazine*, vol. 22, no. 2, pp. 25-37

Freund, Y. and Schapire, R.E. (1996). Experiments with a new boosting algorithm. *Proc. ICML*, pp.148-156.

Gao, Y., Fan, J., Luo, H., Xue, X. & Jain, R. (2006). Automatic Image Annotation by Incorporating Feature Hierarchy and Boosting to Scale up SVM Classifiers, *ACM Multimedia,* Santa Barbara, CA.

Gatica-Perez, D., Lathoud, G. Odobez, J.-M. & McCowan, I. (2007). Audiovisual probabilistic tracking of multiple speakers in meetings. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2), 601 – 616. Feb.

Guironnet, M., Pellerin, D. and Rombaut, M. (2005). Video classification based on low-level feature fusion model. *In: The 13th European Signal Processing Conference*. Antalya, Turkey.

Gupte, S., Masoud, O., Matrin, R. F. K. & Papanikolopoulos, N. P. (2002). Detection and Classification of Vehicles, *IEEE Transactions on Intelligent Transportation System*, vol. 3, no. 1. Pp. 37-47, March.

Hall, D. L. & Llinas, J. (2001). Multisensor data fusion. *In D. L. Hall and J. Llinas (Eds.), Handbook of Multisensor Data Fusion*, CRC Press, pp. 1-10.

Harma, A., McKinney, M., and Skowronek, J. (2005). Automatic surveillance of the acoustic activity in our living environment. *Proc. of the IEEE Int. Conf. on Multimedia and Expo* (ICME)

Heo, J., Kong, S., Abidi, B., & Abidi, M. (2004). Fusion of visual and thermal signatures with eyeglass removal for robust face recognition. *In Proceedings of the Joint IEEE International Workshop on Object Tracking and Classification Beyond the Visible Spectrum*, pp. 94-99.

Hershey, J., Attias, H., Jojic, N., & Krisjianson, T. (2004). Audio visual graphical models for speech processing. *In IEEE International Conference on Speech, Acoustics, and Signal Processing (ICASSP04)*, Montreal, Canada, May.

Hsu, W. L., Yu, S. H., Chen, Y. S. & Hu, W. F. (2006). An Automatic Traffic Surveillance System for Vehicle Tracking and Classification, *IEEE Trans. on Intelligent Transportation Systems*, vol. 7, no. 2, 175-187.

Iwasaki, Y. (2008). A Method of Real-time moving vehicle detection for bad environments using infrared thermal images, *Innovations and Advanced Techniques in System, Computer Science and Software Engineering*, K. Elleithy ed., 43-46, Springer.

Kegl. B & Busa-Fekete, R. (2009). Boosting products of base classifiers. *In ICML*, pages 497-504, 2. 5.

Khan, A.Z., Stanbridge, A.B., & Ewins, D.J. (2000). Detecting damage in vibrating structures with a scanning LDV, *Optics and Lasers in Engineering,* 32, 583-592

Knox, M. T. & Mirghafori, N. (2007). Automatic laughter detection using neural networks. *8th Annual Conference of the International Speech Communication Association*, Belgium, August 27-31.

Kong, S. G., Heo, J., Abidi, B. R., Paik, J., & Abidi, M. A. (2005). Recent advances in visual and infrared face recognition – A review. *Comput. Vision Image Understand*. 97, 1, 103-135.

Krotosky, S. J. & Trivedi, M. M. (2008). Person Surveillance using visual and infrared imagery. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 8, August.

Lai, P., et al, (2008). A robust feature selection method for noncontact biometrics based on Laser Doppler Vibrometry, *IEEE Biometrics Symposium*, 65-70

Leykin, A., Ran, Y., & Hammoud, R. (2007). Thermal-visible video fusion for moving target tracking and pedestrian classification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, R., Wang, T., Zhu, Z. & Xiao, W. (2010). Vibration characteristics of various surfaces using an LDV for long-range voice acquisition, *IEEE Sensor Journal*, November

Li, W., Liu, M., Zhu, Z. and Huang, T. S. (2006). LDV remote voice acquisition and enhancement. *In Proc. IEEE Conference on Pattern Recognition*, 4, 262-265

Li, S. Z. (2000). Content-based classification and retrieval of audio using the nearest feature line method. *IEEE Trans. On Speech and Audio Processing*, September.

Li, X., Chen, G., Ji, Q., & Blasch, E. 2008. A non-cooperative long-range biometric system for maritime surveillance. In *Proc. IEEE Conference on Pattern Recognition*, pages 1-4.

Liu, L., Stamos, I., Yu, G., Wolberg, G., & Zokai, S. (2006). Multiview geometry for texture mapping 2D images onto 3D range data. *IEEE International Conference of Computer Vision and Pattern Recognition*, vol. 2, pp. 2293-2300, June.

Liu, L. & Stamos, I. (2007). A systematic approach for 2D-image to 3D-range registration in urban environments. *VRML Workshop, 11th International Conference on Computer Vision, Brazil*, 14-20, October

Lu, L., Zhang, H-J. & Jiang, H. (2002). Content analysis for audio classification and segmentation. *IEEE Trans. On Speech and Audio Processing*, vol 10, no. 7, October.

Lowe, D. G. (2001). Local feature view clustering for 3D object recognition. *In Proceedings of the Conference of Computer Vision and Pattern Recognition*, Kauai, Hawaii, USA, pp 682-688, December.

Maganti, H. K., Gatica-Perez, D., & McCowan, I. (2007). Speech enhancement and recognition in meetings with an audio-visual sensor array. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8), 2257 – 2269, Nov.

Mao, L., Xie, M., Huang, Y. & Zhang, Y. (2010). Preceding vehicle detection using histograms of oriented gradients. Int. Conf. on Communications, Circuits and Systems (ICCCAS). pp. 354-358, July.

Masagutov, V., Stouch, D. W., Kanjilal, P. & Snorrason, M. (2007). Vibrometry Classification of Moving Vehicles Using Throttle Signature Analysis, *ICIP*.

Nedgård, I. (2005). A comparison of analysis methods for vehicle classification by laser vibrometry, *Swedish Defense Research Agency*, April.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H. & Ng, A. Y. (2011). Multimodal deep learning. *In Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA.

Mikolajczyk, K. & Schmid, C. (2002). An affine invariant interest point detector. *In Proceedings of the 7th European Conference on Computer Vision*, Copenhagen, Denmark, vol. I, pp. 128-142, May.

Mikolajczyk, K, Schmid, C. & Zisserman, A. (2004) Human detection based on a probabilistic assembly of robust part detectors. *In Proceedings of the 8th European Conference on Computer Vision*, Prague, Czech Republic, vol I, pp 69-81.

Mohottala, S., Ono, S., Kagesawa, M., & Ikeuchi, K. (2009). Fusion of a camera and a laser range sensor for vehicle recognition. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, June 20-25, pp. 16-23

Ometron. (2010). Ometron Systems. http://www.imageautomation.com/. Last visited December, 2010.

Petsatodis,T., Pnevmatikavakis, A. & Boukis, C. (2009). Voice activity detection using audio-visual information, *16th Int. Conf. Digital Signal Processing*, August.

Plapous, C., Marro, C., Scalart, P. (2006). Improved signal-to-noise ratio estimation for speech enhancement. *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, issue. 6, pp. 2098-2108, Nov.

Polytec. (2010). Polytec Laser Vibrometer, http://www.polytec.com/. Last visited December, 2010.

Potamianos, G., Neti, C., Gravier, G., Garg, A., & Senior, A. W. (2003). Recent advances in the automatic recognition of audiovisual speech. *Proc. IEEE*, vol. 91, no. 9, pp. 1306-1326, September

Potamianos, G., Neti, C., Luettin, J., & Matthews, I. (2004). Audio-visual automatic speech recognition: An overview. *In Issues in Visual and Audio-Visual Speech Processing*. MIT Press, 2004.

Qu, Y., Wang, T. & Zhu, Z. (2010) vision-aided laser Doppler vibrometry for remote automatic voice detection, *IEEE/ASME Transactions on Mechatronics*, 99, 1-10.

Rabiner, L. R. & Juang, B.-H. (1993). Fundamentals of speech recognition. Upper Saddle River, NJ: Prentice-Hall.

Radhakrishnan, R., Divakaran, A., & Smaragdis, P. (2005). Audio analysis for surveillance applications. *IEEE Workshop on Applications of Signals Processing to Audio and Acoustics*, October 16-19, New Paltz, NY.

Radhakrishnan, R. and Divakaran, A. (2006). Generative process tracking for audio analysis. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP' 06)*, vol. 5, May.

Radova V. & Psutka, J. (1997). An approach to speaker identification using multiple classifiers. In Proc. *IEEE Conf. Acoustics, Speech Signal Processing*, Munich, Germany, vol. 2, pp. 1135-1138.

Redpath, D.B. & Lebart, K. (2005). Boosting feature selection, *Pattern Recognition and Data Mining, Lecture Notes in Computer Science*, Volume 3686/2005, 305-314, DOI: 10.1007/11551188_33

Ronfard, R., Schmid, C. & Triggs, B. (2002). Learning to parse pictures of people. *In Proceedings of the 7th European Conference on Computer Vision*, Copenhagen, Denmark, vol. IV, pp 700-714.

Saberian, M. and Vasconcelos, N. (2012). Boosting algorithms for simultaneous feature extraction and selection. *In CVPR.*

Samadi, S., Kazemi, F. M., Mohamad, R., & Akbarzadeh, T. (2008). Vehicle detection using a multi-agent vision-based system, *Advances in Computer and Information Sciences and Engineering*, T. Sobh ed., 147-152, Springer.

Sanderson, C. & Paliwal, K. K. (2004). Identity verification using speech and face information. *Digital Signal Processing*, vol. 14, no. 5, pp. 449-480

Scotti, G., Marcenaro, L., Coelho, C., Selvaggi, F., & Regazzoni, C. S. (2005). Dual camera intelligent sensor for high definition 360 degrees surveillance. *Vision, Image, Signal Process*. 152, 2, 250-257

Scruby, C. B. and Drain, L. E. (1990). Laser Ultrasonics Technologies and Applications. Madison Avenue, New York: Taylor & Francis.

Seitz, S. & Kim, J. (2003). Multiperspective Imaging. *IEEE CGA*, 23(6), 16-19, 2003

Seo, N. (2007). A comparison of multi-class support vector machine methods for face recognition, *Report*, Dec.

Shah, P., Merchant, S. N., & Desai, U. B. (2010) Fusion of surveillance images in infrared and visible band using curvelet, wavelet and wavelet packet transform. *International Journal of Wavelets, Multiresolution and Information Processing (IJWMIP)*, Volume: 8, Issue: 2(2010) pp. 271-292, DOI: 10.1142/S0219691310003444

Shin, J., Kim, S., Kang, S., Lee, S., Paik, J., Abidi, B., & Abidi, M. (2005). Optical flow-based real-time object tracking using non-prior training active feature mode. *ELSEVIER Real-Time Imaging*, vol. 11, pp. 204-218

Sigurdsson, S., Petersen, K. B. & Lehn-Schioler, T. (2006). Mel frequency cepstral coefficients: an evaluation of robustness of MP3 encoded music. *In Proceedings of the International Symposium on Music Information Retrieval*.

Stauffer, C., Grimson, W.E.L. (1999). Adaptive background mixture models for real-time tracking, *CVPR*, June.

Stomas, I. & Allen P. K. (2000). 3-D model construction using range and image data. *In Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 531-536, June

Sun, Z., Bebis, G. and Miller, R. (2003). Boosting object detection using feature selection. *IEEE International Conference on Advanced Video and Signal Based Surveillance* (AVSS'03).

Tax, D.M.J. and Duin, R.P.W. (2002). Using two-class classifiers for multiclass classification. *Proceedings of Int. Conf. on Pattern Recognition*, Quebec City, Canada, August.

Theodoridis, S. & Koutroumbas, K. (2008). Pattern recognition. 4th ed. Elsevier Science & Technology Books, October.

Thieme, M. J. (2007). Multimodal biometric systems: Applications and usage scenarios. *In Z. Zhu and T. S. Huang (Eds.), Multimodal Surveillance: Sensors, Algorithms, and Systems*, pp. 363-385. Norwood, MA: Artech House.

Tian, Y., Senior, A. W., Hampapur, A., Brown, L., Shu, C., & Lu, M. (2008). IBM smart surveillance system (S3): event based video surveillance system with an open and extensible framework. *Machine Vision and Applications*.

Tian, Y., Feris, R., & Hampapur, A. (2008b). Real-time detection of abandoned and removed objects in complex environments. *The 8th Int'l Workshop on Visual Surveillance (VS).*

Torresan H., Turgeon, B., Ibarra-Castanedo, C., Hebert, P., & Maldague, X. (2004). Advanced surveillance systems: combining video and thermal imagery for pedestrian detection. *In Proc. of SPIE, Thermosense XXVI*, vol. 5405 of SPIE, pp. 506-515, April

Trucco E. & Verri, A. (1998) Introductory Techniques for 3-D Computer Vision. Prentice Hall.

Tsochantaridis, I., Hofmann, T., Joachims, T. & Altun, Y. (2004). Support vector learning for interdependent and structured output spaces, *ICML*.

Vidal-Naquet, M. & Ullman, S. (2003) Object recognition with informative features and linear classification. *In Proceedings of the 9th International Conference on Computer Vision*, Nice, France, pp 281-288.

Vu, V.-T., Bremond, F., Davini, G., Thonnat, M., Pham, Q-C., Allezard, N., et al. (2006). Audio-video event recognition system for public transport security. *Image for Crime Detection and Prevention* (ICDP 2006), London, UK, June.

Wall, M.E., Rechtsteiner, A., Rocha, L.M. (2003). Singular Value Decomposition and Principal Component Analysis. Chap. 5, pp. 91-109, Kluwel, Norwell, MA.

Wang, T., Li, R., Zhu, Z., & Qu, Y. (2011a). Active stereo vision for improving long range hearing using a laser Doppler vibrometer, *IEEE Computer Society's Workshop on Applications of Computer Vision (WACV)*

Wang, T. & Zhu, Z. (2012a) Real Time Moving Vehicle Detection and Reconstruction for Improving Classification, IEEE Computer Society's Workshop on Applications of Computer Vision (WACV), Jan.

Wang, T. & Zhu, Z. (2012b) Multimodal and Multi-task Audio-Visual Vehicle Detection and Classification, 9th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS), Sept.

Wang, T., Zhu, Z. & Blasch, E. (2010a). Bio-inspired adaptive hyperspectral imaging for target tracking*, IEEE Sensors Journal, Special issue on Enhancement Algorithms, Methodologies & Technology for Spectral Sensing*, vol. 10, no 3, pg. 647-654.

Wang, T., Zhu, Z. & Divakaran, A., (2010b). Long-rang, audio and audio-visual event detection using a laser Doppler vibrometer, *SPIE Defense, Security and Sensing: Evolutionary and Bio-Inspired Computation: Theory and Applications IV*, April

Wang, T., Zhu, Z. & Taylor, C. N. (2011b). Real time moving vehicle detection and reconstruction for improving classification. *International Workshop on Video Panorama (IWVP)*, Dec.

Wang, Z. R., Jia, Y. L., Huang, H. & Tang, S. M. (2008). Pedestrian detection using boosted HOG features. *Proceedings of the 11th International IEEE Conference on Intelligent Transportation Systems*, Beijing, China. October.

Weston, J. & Watkins, C. (1998). Multi-class support vector machines, Technical report CSD-TR-98-04.

Xiao, J., Cheng, Hui., Han, F., & Sawhney, H. S. (2008). Geo-spatial aerial video processing for scene understanding and object tracking. *IEEE Conference on Computer Vision and Pattern Recognition*. June.

Yao, Y., Abidi, B., & Abidi, M. (2006). Fusion on omnidirectional and PTZ cameras for accurate cooperative tracking. *In Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS'06)*. Australia, Nov. 22-24

Zhang, C., Yin, P., Rui, Y., Cutler, R. & Viola, P. (2006). Boosting-based multimodal speaker detection for distributed meetings. *In Proc. of IEEE international workshop on Multimedia and Signal Processing*, Oct. Victoria, BC, Canada.

Zhao, J. & Cheung, S. S. (2009). Human segmentation by fusing visible-light and thermal imaginary. *IEEE International Workshop on Visual Surveillance* (ICCV workshop).

Zheng, J. Y. & Wang, X. (2005). Pervasive views: Area exploration and guidance using extended image media. ACM Multimedia Conference, 05, 986-995, Singapore.

Zheng, J. Y. & Tsuji, S. (1990). Panoramic representation of scenes for route understanding. *In Proc. 10-ICPR, IAPR*, June 1990 161-167

Zheng, J. Y., Zhou, Y. & Mill, P. (2006). Scanning scene tunnel for city traversing. *IEEE Trans. Visualization and Computer Graphics*, 12(2), 155-167, 2006

Zhu, Z. & Li, W. (2004). Integration of laser vibrometry with infrared video for multimedia surveillance display, City College of New York, AFRL/HECB Grant Final Performance Report, Dec.

Zhu, Z., Li, W., Molina, E., & Wolberg, G. (2007). LDV sensing and processing for remote hearing in a multimodal surveillance system. *In Z. Zhu and T. S. Huang (Eds.), Multimodal Surveillance: Sensors, Algorithms, and Systems*. pp. 363-385. Norwood, MA: Artech House

Zhu, Z., & Huang, T. S. (eds). (2007). *Multimodal Surveillance: Sensors, Algorithms and Systems,* ISBN-10: 1596931841, Artech House Publisher, July.

Zhu, Z., Xu, G., Yang, B., Shi, D. & Lin, X. (2000). VISATRAM: A real-time vision system for automatic traffic monitoring. *J. Image and vision computing*, 18(10), July, pp. 781-794.

Zieger, C., Brutti, A. and Svaizer, P. (2009). Acoustic based surveillance system for intrusion detection. IEEE ICVSBS'09: 314-319.

Zivkovic, Z. (2004). Improved adaptive Gaussian mixture model for background subtraction. *ICPR*, UK, Aug.

Zivkovic, Z. and van der Heijden F. (2006). Efficient adaptive density estimation per image pixel for the task of background subtraction, *Pattern Recognition Letters*, vol. 27, no. 7, pp. 773-780.

Zotkin, D. N., Duraiswami, R. & Davis L. S. (2001). Multimodal 3-D tracking and event detection via the particle filter. *Proc. of the IEEE Workshop on Detection and*

Recognition of Events in Video (in association with ICCV 2001), Vancouver, Canada, pp. 20-27

Zotkin, D. N., Raykar, V. C., Duraiswami, R., & Davis, L. S. (2007). Multimodal tracking for smart videoconferencing and video surveillance. In Z. Zhu and T. S. Huang (Eds.), *Multimodal Surveillance: Sensors, Algorithms, and Systems.* pp. 141-175, Norwood, MA: Artech House, 2007.

Zou, X. & Bhanu, B. (2005). Tracking humans using multimodal fusion. *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, US.