8-1-2012

Wavelets in Intelligent Transportation Systems: Data Compression and Incident Detection

Shaurya Agarwal
University of Nevada, Las Vegas

WAVELETS IN INTELLIGENT TRANSPORTATION SYSTEMS: DATA COMPRESSION AND INCIDENT DETECTION

by

Shaurya Agarwal

Bachelor of Technology
Indian Institute of Technology, Guwahati
2009

A thesis submitted in partial fulfillment
of the requirements for the

Master of Science in Electrical Engineering

Department of Electrical and Computer Engineering
Howard R. Hughes College of Engineering
The Graduate College

University of Nevada, Las Vegas
August 2012

UNLV
UNIVERSITY OF NEVADA LAS VEGAS

THE GRADUATE COLLEGE

We recommend the thesis prepared under our supervision by

ShauryaAgarwal

entitled

Wavelets in Intelligent Transportation Systems: Data Compression and Incident Detection

be accepted in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering
Department of Electrical and Computer Engineering

Pushkin Kachroo, Committee Co-Chair

Emma Regentova, Committee Co-Chair

Ke-Xun Sun, Committee Member

Haroon Stephen, Graduate College Representative

Thomas Piechota, Interim Vice President for Research and Graduate Studies
and Dean of the Graduate College

August 2012

ABSTRACT

WAVELETS IN INTELLIGENT TRANSPORTATION SYSTEMS:

DATA COMPRESSION AND INCIDENT DETECTION

by

Shaurya Agarwal

Dr. Pushkin Kachroo, Examination Committee Co-chair
Professor, Electrical and Computer Engineering Department
University of Nevada, Las Vegas

Dr. Emma Regentova, Examination Committee Co-chair
Associate Professor, Electrical and Computer Engineering Department
University of Nevada, Las Vegas

Research show that wavelets can be used efficiently in denoising and feature extraction of a given signal. This thesis discusses about intelligent transportation systems(ITS), its requirement and benefits. We explore use of wavelets in intelligent transportation systems for knowledge discovery, compression and incident detection. In the first section of thesis, we focus on the following problems related to traffic matrix: data compression, retrieval and visualization. We propose a methodology using wavelet transform for data visualization and compression of traffic data. Aim is to research on the wavelet compression technique for the traffic data, come up with the performance of various available wavelets and the best decomposition level in terms of

compression ratio and data distortion. We further investigate use of Embedded Zero Tree (EZW) encoding and Set Partitioning in Hierarchical Trees (SPIHT) algorithm for compression of the traffic data.

In the second section of thesis, we focus on regression model for dichotomous data, i.e. logistic regression. This model is suitable when the outcome can takes only limited number of values, in our case only two, presence or absence of an incident. We look into generalized linear model (gle) with binomial response and logit link function. We present a framework to use logistic regression for incident prediction in transportation systems. Further in the section, we investigate feature extraction using DWT, and effect of preprocessing of data on the performance of incident detection models. A hybrid logistic regression-wavelet model is proposed for traffic incident detection.

ACKNOWLEDGMENTS

I would like to express my sincere thanks and gratitude to my supervisor, Dr.Pushkin Kachroo, who has supported and motivated me throughout my thesis. He continuously instilled a spirit of adventure and excitement towards research and learning in me. He nurtured the thesis throughout with patience and motivation at the same time allowing me to explore things in my own way. I simply could not wish for a better or friendlier supervisor.

I am highly grateful to my co-supervisor, Dr.Emma Regentova for her guidance and motivation. She has been really helpful and kind throughout, always ready to share knowledge and enlighten us with all her experience and knowledge.

I am also grateful to my committee members, Dr.Ke-Xun Sun and Dr.Haroon Stephen for their guidance and insightful remarks.

I am also thankful to friendly and cheerful group of fellow students namely Puneet, Pratik, Atul, Romesh, Sourabh, Himanshu, Shivam, Anuj and Pankaj for all their support and help during the period of my work.

Finally, I would like to take this opportunity to show my appreciation towards the love of my parents Mr.Sanjay Agarwal and Dr.Mohini Goel who have been my source of inspiration since my childhood.

Dedication

"Guru Govind dou khade, kaake laagoon paye

Balihari guru aapne, Govind diyo milaye."

- Sant Kabir

Translation: If my Teacher and my God, are present in front of me at the same time, I will offer the respect, first to my Teacher, as he is the one, who showed me the way to God (Light).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Background

Leading cause of death among those between ages 5-34 in the U.S. is motor vehicle crashes. For the age group from 1 to 45 years of age, on an average, 47.5% of deaths are caused by moving traffic in 2007 [1]. More than 2.3 million adult drivers and passengers were treated in emergency departments as the result of being injured in motor vehicle crashes in 2009 [2]. $70 billion is attributed to costs associated with crash-related deaths and injuries among drivers and passengers in 2005 (2).

Traffic accidents often result in fatalities, severe injuries, mental traumas, pain, tremendous medical bills and insurance premiums. These road crashes take a high toll from citizens physically, economically and mentally. Seat belts are one of the most effective passive safety features in vehicles and there is a host of research literature attesting to the effectiveness of seat belts in protecting against death and injury. Clearly, even when use rates are high the potential gains in trauma reduction from further improvements in wearing rates are substantial. However, those currently most resistant to restraint use have also proven most difficult to target using conventional countermeasures. In Nevada, the distribution of the population and fatal crashes has changed over the years. Although statistics were available every year to speak of the

traffic fatality trends, the crash and medical datasets available could not be brought to a common platform to be studied together. Later, an integrated repository on traffic related injuries, deaths, crash rates and medical information was prepared by Center for Traffic Safety Research, joining the NDOT (Nevada department of traffic) crash data and UMC trauma data from 2003-2008. In addition to this other transportation data includes: Flow Detector Data, AVL data and Bluetooth data received from RTC. UNLV also collects its own traffic related data, such as yearly seatbelt survey, road construction data and data collected through smart phone applications.

1.2 Motivation

Huge amount of traffic data has created a need for so called Intelligent Transportation System (ITS), which is able in storing, retrieving and analyzing the data in an efficient and optimal way. The data obtained through various sensors and surveys are in the form of large spatio-temporal series. Challenge before the ITS is the huge amount of traffic data, their storage, retrieval and transmission. Traffic research needs this data for pattern matching, data mining, database queries, prediction and trend analysis. Efficient data compression and data retrieval techniques are the two main desired features of ITS. Main issues with the traditional data processing techniques are the compression efficiency, data loss and redundancy.

Transportation system has multiple datasets that are often maintained separately, such as roadway data, intersection data, crash data, road construction data etc. It is very important that all the information in these datasets is linked together to

allow analysis with real time updates. Hence an integrated, intelligently designed and optimized relational database is a must for managing any transportation system. The system should create a common integrated data base for the better coordination between the police, maintenance and road construction department and the traffic and crash management authorities.

1.3 Research Goal

Research show that wavelets can be used efficiently in denoising and feature extraction of a given signal. We explore use of wavelets in intelligent transportation systems for knowledge discovery, data management, compression and incident detection. Wavelet transform allows us to separate common patterns throughout the traffic data and deviations from the average flow by capturing differences from the average flow. Additionally, wavelet multiresolutional analysis allows for convenient way of representing data and retrieval of the data of interest directly from compressed data. Similarities and differences in traffic flow can be found by analysis of wavelet coefficients, that is approximation and the details. This thesis mainly addresses the issues and challenges faced for an efficient management of transportation data and try to find the answers using wavelets. Study is divided into two parts:

First section is about knowledge discovery, data management and compression techniques using wavelets. Traffic Matrix provides amount of traffic flow between each origin-destination pair in a network. Traffic matrices contain huge amount of data as they are collected for a long period at short intervals. It is often observed

that data is repetitive and follows a spatio-temporal pattern. For example consider traffic flow data which is expected to be more or less same at a particular point during peak hours on weekdays. It might be different on weekends or other hours of the day. Hence instead of storing whole data there is a need to capture the essence of the data. In this section we focus on the following problems related to traffic matrix: data compression, retrieval, visualization and prediction. Wavelet compression is a form of data compression technique well known for image compression. The traffic data obtained by various sensors are mainly in numeric format that can be treated as a signal, and hence the wavelet compression techniques can be applied. As the wavelet compression results in the lossy output, there is a tradeoff between the data distortion and the compression ratio.

Second section explores incident detection algorithm for traffic incident detection. We investigate logistic regression for incident detection. Logistic Regression is type of regression analysis where we can predict categorical outcomes based on certain predictors. Probabilities of the possible outcomes are modeled, using logistic functions, as a function of independent variables. We further investigate use of DWT for preprocessing the data for improvement in incident detection. A hybrid logistic regression-wavelet model is proposed for traffic incident detection. This model uses DWT for filtering and denoising of the traffic data before applying logistic regression for incident detection.

CHAPTER 2

INTELLIGENT TRANSPORTATION SYSTEMS

2.1 Introduction

A spatially enabled traffic management system is a must for a local transportation authority for a better management, policy framing and mitigation of traffic incidents. The main components of the system must include a comprehensive integrated database, inbuilt statistical tools, data visualization on maps and report generation ability. Apart from the use of statistics and graphical representation, spatial visualization of the data and maps displaying the traffic patterns and collisions are a must for a clear and better understanding. Use of GIS can transform the tabular data into a more meaningful presentation. An ideal system must include features such as Crash Prediction, Policy Review, Design Consistency, Traffic Analysis, Driver/Vehicle and Intersection Review. Some of the main objectives of the Safety Analysis System are as follows [3]:

- To be able to retrieve and visualize crash data from the system
- To be able to identify most crash prone road segments and intersections
- To be able to retrieve historical crash data for statistical analysis and traffic studies

- To be able to identify correlations between collision types and their causing factors
- To be able to update the system and database periodically with the new crash data
- To be able to provide spatial database of road network including information about pavements, location of traffic control devices including stop signs and other signals, traffic volume, posted speed limits and actual speed limits.
- To be able to generate comprehensive reports by ranking risk prone roadway segments, collision factors and conditions.
- To be able to develop an easy to use interface that can be used by general public for getting to know about traffic pattern and crash prone intersections.

To understand better we researched on what software are being used for safety analysis purposes by different states across the country [4]. Findings are listed below in table 2.1.

State	*SafetyAnalysisSoftware*
California	Traffic Accident Surveillance, Transportation System Network
Florida	Crash Analysis Reporting System (CARS)
Georgia	Accident Information System (AIS), CARE, Safety Analyst
Idaho	WebCARS
Iowa	Crash Mapping Analysis Tool (CMAT), SAVER, Crash Magic
Kansas	High Accident Location System (HAL), Safety Analyst
Michigan	Crash Processing System (CPS), Safety Management System
Missouri	Safety Management System, Safety Analyst
Nebraska	Hazardous Location Program
New Hampshire	Safety Analyst
Ohio	Ohio Crash Location Analysis Tool
Pennsylvania	Crash Data Analysis and Retrieval Tool (CDART)
Virginia	Road Network System

Table 2.1: Safety Software Used by Various States

2.2 Functional Requirement Specification for ITS

1. **Integrated Relational Database:**

   Transportation system has multiple datasets that are often maintained separately, such as roadway data, intersection data, crash data, road construction data etc. It is very important that all the information in these datasets is linked together to allow analysis with real time updates. Hence an integrated, intelligently designed and optimized relational database is a must for managing any transportation system. The system should create a common integrated data base for the better coordination between the police, maintenance and road construction department and the traffic and crash management authorities.

2. **GIS Integration:**

   System should have the ability to geo-reference all accident data on a digital road map and to instantly analyze the information by means of the powerful

database. GIS analysis is important due to the ease of its use and due to all the visual tools that come along with it. GIS is considered as an important tool in managing, planning, evaluating, and maintaining transportation systems.

3. **Linear Referencing System:**

   Linear referencing system is a convenient means to associate attributes or events to locations or segments of a linear feature. This ability to associate attributes to a point or segment is very important from GIS as well as data storage point of view. Data storage needs can be reduced if the future updates on collisions or attributes are recorded spatially. LRS is widely used in transportation applications.

4. **Statistical Analysis:**

   System should be able to perform various statistical tests on the datasets and attributes we choose. Examples of this include crash locations, conditions, crash frequency and crash rate, all of which help to relate crash data into meaningful statistics for comparison of locations and the factors. This analysis must be coupled with good plotting tools for viewing the results in various formats such as pie charts and bar graphs for quick and better understanding.

5. **Network Analysis:**

   Transportation Network Analysis is concerned primarily with the spatial, but also the temporal, movement of people in a region through roadway networks or other means of transportation. Apart from location specific analysis, system

should also be able to view the network as a whole and perform analysis. This could help in better understanding of the system and can help to pin point the areas where improvements are needed in the network.

6. **Report Generation Capability:**

   The system should have real-time reporting and downloading features for keeping the authorities and general public up-to-date. Further it should have the ability to store historical data for previous years, hence ensuring data mining capabilities. The system should not only help with the current operations but also have the ability to analyze future trends and patterns. The ability to analyze data and summarize it into a report directly from the integrated crash database is desirable. For the ease of use, standard report formats should be inbuilt into the system, at the same time system should be flexible enough to provide the option of customized reports as per needs.

7. **Periodic Database Updation:**

   Data for evaluating system performance is collected regularly by transportation agencies including information related to new crashes, road conditions, pavement conditions, travel times, crash rates, etc. This data must be updated into the system with precision and in a timely manner to keep the system up to date.

8. **Integrated System:**

   An integrated management system is a system that integrates all of an orga-

nization's systems and processes in to one complete framework, enabling an organization to work as a single unit with unified objectives. The Freeway Operations Committee of the Transportation Research Board has defined ITMS as: An integrated transportation management system (ITMS) provides for the automated, real-time sharing of information between ITS based systems and the coordination of management activities between transportation agencies, thereby enhancing system interoperability and enabling an area wide view of the transportation network. These systems and agencies provide for the management and operation of a variety of different transportation facilities and functions, including freeways, arterial streets, transit (bus and rail), toll facilities (e.g., bridges, tunnels), emergency service providers, and information service providers. [5]

2.3 Conclusion

Figure 2.1 shows how tasks of a modern day transportation agency revolve around an ITS.

ITS represents a wide collection of applications, from adaptive traffic signal control systems, to automated incident/congestion warning system, to intelligent ramp meters, to collision avoidance systems, to mode of creating public awareness. The ultimate benefits of an integrated transportation system are of wide spectrum making it a fully connected; information-rich; safe, efficient, and environment friendly. ITS enables significant improvement in transportation system performance, including reduced congestion and increased traveler convenience and safety. ITS can be very

Figure 2.1: ITS

useful in helping any transportation agencies in their current objective and functions as well as for achieving futuristic goals.

Some of the major functions of any transportation agency (e.g. NDOT and RTC in Nevada) include controlling signalized intersections and ramp meters, managing Dynamic Messaging Signs, alerting citizens of traffic incidents through sms and emails, creating public awareness etc. The below list describes how these agencies can use ITS to perform these functions more efficiently and effectively:

1. Real-time Traffic Information System

2. Automatic Incident Reporting Mechanism
3. Adaptive Traffic Signal Control
4. Automatic Dynamic Message Signs
5. Intelligent Real-time Ramp Metering
6. Real-time Status Information for Public Transit System
7. Parking Information
8. Navigation Systems
9. Weather Information Systems
10. Integrated Real-time Interactive Dashboard
11. Automatic Sms and email alerts to subscribers

CHAPTER 3

DATA SOURCES AND CHARACTERISTICS

Following is the list of major sources of various kinds of traffic and transportation related data that TRC acquires periodically.

- FAST- Freeway and Arterial System of Transportation
- LVMPD- Las Vegas Metropolitan Police Department
- NDOT- Nevada Department of Transportation
- NHP- Nevada Highway Patrol
- TRC, UNLV- Transportation Research Center, University of Nevada Las Vegas.
- UMC- University Medical Center

3.1 Data Sources and Characteristics

Following is the list and brief description of data available with TRC at present:

3.1.1 FAST

1. Flow Detector Data: The flow detector data is available to TRC as a live feed for two highway stretches at present; US-95 and I-15. It records the data every

minute about the occupancy of various lanes. Parameters that are available in the flow detector data are -

- Date and time stamps
- Detector IDs
- Lane wise vehicle count
- Occupancy
- Lane speed

2. SMS Data

- Data and time
- Location
- Lane Blocked

3. AVL Data: The AVL data is recorded on the transit vehicles under Regional Transportation Commission (RTC) of Southern Nevada. The current data that is available is for routes 110 and 202. The route 110 (Eastern) functions between Cheyenne/Civic Center and Eastern/St. Rose Pkwy; while the route 202 (Flamingo) functions between Fort Apache/Flamingo and Harmon/Boulder Hwy. With the sensors in the vehicles, the time of arrival of the vehicle at various stops is noted and compared with the stipulated time of arrival at the stop. This allows calculating the delays at various stops which are saved in a data file. The parameters that are recorded into a data file are listed below -

- Data and time
- Coach Number
- Block Number
- Trip Number
- Stop Locations (Names)
- Arrival Time at the stops
- Delay calculated from the stipulated time of arrival

4. Bluetooth Data: The Bluetooth data is collected by paired bluetooth detectors that are installed on the roads. The current paired bluetooth sensors are located between Hwy 93 & Lakeview Dr (u182) and Hwy 93 & West of Veterans Memorials Dr (u179). The bluetooth data is recording by getting the ID of any available Bluetooth device in a vehicle which crosses the sensor at a location. At the location of the next sensor, the ID is matched to calculate the time taken by the vehicle to traverse the distance between the sensors. Since the distance between the sensors is known, it can be used to calculate the speed of the vehicle. The data is publicly available on bluetoad.trafficcast.com. The data collected is filtered and has the following parameters -

   - Calculated Time
   - Last match Time
   - Travel Time

- Speed

3.1.2 LVMPD

1. Arterial Incident Management (IM) Data

    - Event Number
    - Create Time
    - Arrival Primary Unit
    - Cleared Time
    - Code

3.1.3 NDOT

1. Crash Data: The accident data is compiled by NDOT, which includes very detailed information about the recorded crashes. The data covers various aspects ranging from the location and time of the accident to the number of fatalities and roadways conditions. A few major parameters that are recorded in the crash data are listed below

    - Date and Time stamp
    - Type of Accident (Hit and run/ vehicle collision)
    - Collision Description
    - Crash Time Origin and Clearance

- Type of Damage
- Distance from Street
- Roadway type and number of lanes
- Lighting Conditions at time of accident
- Number of fatalities and injuries

Crash data is collected by various agencies like Las Vegas Metropolitan Police Department (LVMPD) for Las Vegas, National Highway Patrol (NHP), Sheriff Offices and other sources. Department of Motor Vehicles (DMV) collects data from all these sources and passes on to Nevada Department of Traffic (NDOT) which compiles the crash data happening around the Clark County. Office of Transportation Safety (OTS) Nevada along with University Medical Centre (UMC) Las Vegas and NDOT worked to link the NDOT crash data with the UMC Trauma data. This data is broken into 6 tables each covering an aspect of the crash like vehicle information, location information, crash information, etc. Some of these accidents generate a trauma support request i.e. a 911 call for emergency is made to support victims. It is this data which gets surfaced in trauma records of hospital. Using this information as the formal basis, both the sets of data were linked based on unique accident identification. The received data was then de-identified by UMC; that is all the personal information (like name and address) was removed before being handed over for analysis to TRC, UNLV. The linked data comprised of UMC trauma data for the period 2005-09

and NDOT crash data for 2005-08. Many hospitals in Clark County having the trauma department contributed with the data. Therefore not all the Clark county data is available. NDOT crash data was available for many accidents but since trauma data was the limiting factor, appropriate records were selected from it and linked to crash data. The final linked crash-trauma dataset had 4112 records for the period of 2005-08.

3.1.4 NHP

1. Freeway Incident Management (IM) Data

    - Date and time
    - Location
    - Place
    - Type Of Accident
    - Receive Time
    - Dispatch Time

3.1.5 TRC-UNLV

1. Construction Data

    - Date and time
    - Location

- Closure point
- End point
- Detour
- Scheduled Work

2. Seatbelt Data: Seatbelt data is collected by TRC every year as part of the statewide seatbelt usage surveys. The data is collected manually through data collection software which was designed in TRC on a PDA (Personal Digital Assistant). According to the uniform criterion the study is conducted in two counties i.e. Clark and Washoe, selected based on population. There are 32 sites each in both the counties where data is collected. The parameters collected during the study are seatbelt status of the front seat occupants, age group, ethnicity, gender, vehicle type and license of registration. Survey provides us with an unweighted estimate of seatbelt usage. It has the following attributes:

   - Gender
   - Ethnicity
   - Age
   - Vehicle category
   - State of registration

3. Travel Run Data

   - Date and time

- Location
- Speed
- Traffic Light Status
- Stopping time at red light
- Proceeding time at light

4. iPhone Application Data The iphone application developed by TRC, UNLV is used to collect data on travel runs. The device can be simply be mounted in the vehicle and it collects the data as you drive around. The iPhone can record the parameters at a pre-programmed time interval, which can be changed to allow the frequency of recording data. The various parameters that are recorded are listed below -

    - Time Stamp
    - Accelerometer data along the three axes
    - Gyroscope data along the three axes
    - Co-ordinates (Latitude and Longitude)
    - Distance interval between recordings
    - Total distance traveled
    - Speed

3.1.6 UMC

1. Trauma Data

    - Accident Information
    - Patient Information
    - Vehicle Information
    - Traffic Conditions
    - Weather Information
    - People Information

3.2 Conclusion

This chapter listed the data sources and characteristics, which TRC manages and analyzes. Is is clear that the amount of data generated and used by ITS is huge. It requires special techniques to store the data efficiently.

CHAPTER 4

INTRODUCTION TO WAVELET ANALYSIS

4.1 Introduction

Fourier transform is a powerful tool to analyze the frequency components of the given signal, but it has some limitations. It cannot tell at what time a particular frequency dominates. Short time frequency Transform (STFT) can be used which uses a sliding window technique to find spectrogram which contains both frequency and time domain information. But this scheme has also a drawback of a limited resolution in frequency domain. Wavelet transform can resolve this problem. Wavelet Transform is a multi-resolution transform which allows a simultaneous time-frequency analysis. Fourier Transform gives a precise analysis of frequencies of the signal but not the time when those frequencies occurred. Wavelet Analysis differs here with the Fourier Analysis. Wavelet Analysis uses basis functions that are local in scale and space where as Fourier analysis uses sine and cosine functions, which are not local in space. Lack of localized support in Fourier transform made them susceptible to Heisenberg's Uncertainty principle. Hence we can say that wavelet transform was born out of need to overcome limitations of Fourier transforms.

4.2 Basic Concepts

4.2.1 What is a Wavelet?

A wavelet is a piece of wave. We can define a wavelet, simply as a function of a short duration in time which has exactly the same area as above and below x-axis. Fourier transforms uses an infinitely repeating sinusoidal wave whereas a wavelet exists only within finite time duration and is zero elsewhere else. A wavelet transform is performed by convolving the signal against particular instances of the wavelet at various time scales and positions. By modeling changes in frequency (by adjusting the time scale) and modeling time changes (by shifting the position of the wavelet), we can model both frequency changes and location of the frequency.

4.2.2 Wavelet Analysis

In the wavelet transform we get information about when certain features occurred and information about the scale characteristics of the signal. Scale can be described as analogous to frequency, and is a measure of the amount of detail present in the given signal. Scale is a number related to the number of coefficients and is counter-intuitive to the level of detail. Small scale generally means gross details, and large scale means fine details. In wavelet analysis we first represent our signal as a linear combination of wavelets. This linear combination is formed by using translations and scalings of only one 'mother wavelet'. Coefficients corresponding to the wavelets indicate the significance of that wavelet. These coefficients are called wavelet transform of the data and they highlight local features of the data. As per needs further processing

of these coefficients can be done, this may include a de-noising feature extraction, clustering or compression problems. Finally, inverse wavelet transform is computed to reconstruct back the data in the original domain.

4.2.3 Types of Wavelets

Many types of wavelets have been developed since 1980s. We can categorize them based on their characteristic properties into following categories:

- Support: Wavelets can be divided into two groups based on their support; compactly supported and infinitely supported. However for infinitely supported wavelets also the function dies down at a very fast rate. Examples of compactly supported wavelets include Haar and Daubechies where as Meyer wavelet is an example of infinitely supported wavelet.

- Smoothness: Wavelets can be differentiated on the basis of their smoothness (differentiability). On one side there is Haar wavelet which is not differentiable even once and on other hand there are wavelets like Meyer, which are continuously differentiable.

- Shape: Wavelets can be symmetric or asymmetric. For example Meyer is a symmetric wavelet whereas Daubechies is an asymmetric wavelet.

Some basic type of wavelets:

1. Haar Wavelet: Haar wavelet is the most basic kind of wavelet. It is a sequence of rescaled "square-shaped" functions which together form a wavelet family or

basis. The Haar sequence was proposed in 1909 by Alfrd Haar. It is considered as a special case of the Daubechies wavelet and also known as D2. It is the simples possible wavelet.

2. Daubechies Wavelet: Ingrid Daubechies invented the compactly supported orthonormal wavelets, thus making discrete wavelet analysis practicable. The names of the Daubechies family wavelets are written dbN, where N is the order, and db the "surname" of the wavelet. The db1 wavelet is the same as Haar wavelet.

4.3 Orthonormal Wavelet Bases

Computation of DWT requires existence of orthogonal bases for $L^2(R)$. Such bases were not available initially in wavelet analysis posing many mathematical problems. It was only after 1980s that mathematicians came up with the methodology of formulating these orthonormal bases. We will introduce Haar bases in this section, a simple orthonormal wavelet bases of L^2, and steps to construct it [11].

4.3.1 Approximation by Step Functions

We know that any L^2 function can be approximated with the help of step functions. Let us consider a step function as described below:

$$\chi_{n,k}(x) = \begin{cases} 1, & 2^{-n}k \leq x < 2^{-n}(k+1) \\ 0, & \text{otherwise} \end{cases}$$

Then for any function $f \epsilon L^2$, there exist step functions

$$f_n(x) = \sum_{k \epsilon \mathbb{Z}} c_{n,k} \chi_{n,k}(x), \qquad n \, \epsilon \, \mathbb{N}$$

such that,

$$\lim_{n \to \infty} \|f_n - f\| = 0$$

Now we define

$$V_n = \left\{ g_n \mid g_n = \sum_{k \epsilon \mathbb{Z}} a_{n,k} \chi_{n,k}(x), \quad a_{n,k} \, \epsilon \, l^2 \right\} \tag{4.1}$$

where V_n is a sequence of subspaces of L^2 which approximates L^2. For every function f in L^2, we have a step function $f_n \, \epsilon \, V_n$ such that

$$\lim_{n \to \infty} \|f_n - f\| = 0$$

It can be observed that as the value of n increases, resolution of V_n increases. Hence the subspaces V_n are nested as described below:

$$... \subset V_{-1} \subset V_0 \subset V_1 \subset ...$$

It is observed that

$$\overline{\bigcup_{n\epsilon\mathbb{Z}} V_n} = L^2 \tag{4.2}$$

and

$$\bigcap_{n\epsilon\mathbb{Z}} V_n = \{0\} \tag{4.3}$$

Now we first construct the orthonormal basis of v_0. We define box function $B(x)$ as:

$$B(x) = \begin{cases} 1, & 0 \leq x < 1 \\ 0, & \text{otherwise} \end{cases}$$

Orthonormal basis of V_0 is formed by $\{B(x-k)\}_{k\epsilon\mathbb{Z}}$. Now orthonormal basis for any V_n can be obtained by dilating $\{B(x-k)\}_{k\epsilon\mathbb{Z}}$.

4.3.2 Haar Wavelet Bases

Now by using equation 4.2, we will try to construct orthonormal basis of L^2. Let W_n be the orthogonal complement of V_n with respect to V_{n+1}

$$W_n \oplus V_n = V_{n+1} \text{ and } W_n \perp V_n \tag{4.4}$$

By using the equations 4.3 and 4.4 we can say that:

$$L^2 = \bigoplus_{n \epsilon \mathbb{Z}} W_n \quad \text{and} \quad W_n \perp W_{n'},\ n \neq n' \tag{4.5}$$

Now as each V_n is a dilation of V_0, W_n is also a dilation of W_0, hence an orthogonal basis of W_0 is also a basis of W_n. We define Haar function as:

$$H(x) = \begin{cases} 1, & 0 \leq x < \frac{1}{2} \\ -1, & \frac{1}{2} \leq x < 1 \\ 0, & \text{otherwise} \end{cases}$$

It can be proved that $H_k(x) = H(x-k)$ is orthonormal basis of W_0 and $H_{n,k}(x)$ is orthonormal basis of W_n.

4.4 Discrete Wavelet Transform (DWT)

Among all the types of wavelet transforms, DWT is most commonly used for discrete signals. Following two sections describe mathematical theory and methodology to compute DWT in detail.

4.4.1 Mathematics of DWT

In this section we will explore mathematics behind discrete wavelet transform [12]. We know that any function $f \epsilon L^2(R)$ can be written as linear combination of

elementary functions $\psi_{j,k}(x)$:

$$f(x) = \sum_{j,k} w_{j,k}\psi_{j,k}(x), \quad j,k \,\epsilon\, \mathbb{Z} \tag{4.6}$$

where $w_{j,k}$ is the set of coefficients. Subscripts j and k are used to indicate the two dimensional decomposition for providing resolution in both time and frequency domain. We obtain elementary function from the 'mother' wavelet in the following manner:

$$\psi_{a,b}(x) = \frac{1}{\sqrt{a}}\psi(\frac{x-b}{a}) \quad a > 0,\ b \,\epsilon\, \mathbb{Z} \tag{4.7}$$

where a and b are integers and represent scaling and translation respectively. As mostly in practical uses scaling is done in powers of two, hence the dyadic version of the above equation can be written as:

$$\psi_{j,k}(x) = 2^{\frac{j}{2}}\psi(2^j x - k), \quad j,\ k \,\epsilon\, \mathbb{Z} \tag{4.8}$$

If the $\psi_{j,k}$ forms an orthonormal basis then the coefficients of DWT can be computed by taking inner product of the function $f(x)$ with the wavelet $\psi_{j,k}$:

$$w_{j,k} = \langle f, \psi_{j,k} \rangle = \int_R f(x)\psi_{j,k}\, dx \tag{4.9}$$

4.4.2 Computing the DWT

The approach discussed in the previous section to compute DWT coefficients by taking the inner product of the basis function ($\psi_{j,k}$) and the original signal ($f(x)$) is computationally extensive. It requires many inner product calculations and hence in general not used for practical reasons.

Mallat came up with an efficient algorithm in 1989, which used multiresolution property of wavelets for computation of DWT. Generally discrete wavelet transform uses the dyadic scheme. To perform the wavelet transform we separate the highpass and lowpass components of the signal by breaking it into a lowpass (scaling function) and a highpass (wavelet function) subbands. We continue the same process downward along the lowpass subband. At each stage, convolution and downsampling are performed to get the input for the next stage.

$$c_{j,k} = c_j[k] = \sum_m h[m - 2k]c_{j+1}[m] \tag{4.10}$$

$$d_{j,k} = d_j[k] = \sum_m h_1[m - 2k]c_{j+1}[m] \tag{4.11}$$

Equation 4.10 and 4.11 give the recursive algorithm for computing DWT. $h[n]$ and $h_1[n]$ represent scaling and wavelet functions respectively. The process described by these two equations can be seen as a convolution process followed by a downsampling. The lowpass signal is treated as an original and is further subdivided into its own low and high subbands. This iteration can be performed all the way down until the lowpass and highpass bands are left to single value, but the stop criterion depends

on the application and the desired output and objective. As this structure looks like a pyramid, hence Mallat's algorithm is also known as pyramid algorithm.

The wavelet and scaling filters must be orthogonal so that the highpass and low-pass information are mutually exclusive. This ensures that there is no overlap between the data representations in frequency domain. To keep the energy of the wavelet constant for every depth we make the integral of the scaling filter $\sqrt{2}$. The Discrete Wavelet Transform can be described as a series of filtering and sub-sampling. The coefficients are calculated by applying a high-pass wavelet filter to the signal followed by down-sampling the resulting signal by a factor of two. At the same level, a low-pass scale filter is also applied followed by down-sampling, producing the signal for the next level.

Scale-coefficients thus obtained are indicative of smoothening of the signal by removal of details, whereas the wavelet-coefficients correspond to the differences between the scales. We can reconstruct the original signal using the sale and wavelet coefficients. The total number of scale coefficients plus wavelet coefficients is equal to the number of samples in the original signal.

CHAPTER 5

VISUALIZATION AND COMPRESSION OF TRAFFIC DATA USING WAVELETS

5.1 Introduction

Huge amount of traffic related sensor data has created a need for so called Intelligent Transportation System (ITS), which is able in storing, retrieving and analyzing the data in an efficient and optimal way. The data obtained through sensors is in the form of large spatio-temporal series. Challenge before the ITS is the huge amount of traffic data, their storage, retrieval and transmission. Traffic research needs this data for pattern matching, data mining, database queries, prediction and trend analysis. Intelligent data compression and data reconstruction techniques are the two main desired features of ITS. Main issues with the traditional data processing techniques are the compression efficiency, data loss and redundancy. Traffic Matrix provides amount of traffic flow between each origin-destination pair in a network. Traffic matrices contain huge amount of data as they are collected for a long period at short intervals. It is often observed that data is repetitive and follows a spatio-temporal pattern. For example consider traffic flow data which is expected to be more or less same at a particular point during peak hours on weekdays. It might be different on weekends or other hours of the day. Hence instead of storing whole data there is a

need to capture the essence of the data.

In this section of thesis, we focus on the following problems related to traffic matrix: data compression, retrieval and visualization. Wavelet transform technique will allow us to separate common patterns throughout the traffic data. Wavelet compression is a form of data compression technique well known for image compression. The traffic data obtained by various sensors are mainly in numeric format that can be treated as a signal, and hence the wavelet compression can be applied. As the wavelet compression results in the lossy output, there is a tradeoff between the data distortion and the compression ratio. Aim is to research on the wavelet compression technique for the traffic data, come up with the performance of various available wavelets and the best decomposition level in terms of compression ratio and data distortion. We propose a schema for data arrangement for compression and knowledge discovery. We further investigate use of Embedded Zero Tree (EZW) encoding and Set Partitioning in Hierarchical Trees (SPIHT) algorithm for compression of the traffic data.

5.2 Problem Statement

In this chapter we will focus on investigating following issues from point of view of transportation needs:

- Significance of Approximate curve (scaling functions) and fine details (wavelet functions) in traffic data.
- Retrieval of features from the reconstructed signal.

- A schema for data arrangement in the file for analysis. e.g. Day wise data from a single detector in a 1D format, or multiple day's data in a single file as a 2D matrix and 3D matrix etc.

- Framework to choose the right level of decomposition and right wavelet as per requirements.

- Analyze the performance of the selected wavelet and decomposition level in terms of compression ratio and signal distortion of the reconstructed signal.

- Data Visualization using color coding of the obtained wavelet decomposition of the signal.

- Investigate Embedded ZeroTree Encoding and SPIHT algorithm as compression techniques for the traffic data.

5.3 Literature Review

Qiao et al, proposed a method which incorporates the one-dimensional discrete wavelet compression approach for Intelligent Transportation Systems Data [23], [24]. As the proposed wavelet compression is a lossy algorithm, balance between the compression ratio and the data distortion is very important. Hence the optimal selection of a threshold is required for balancing of compression parameters. An algorithm is proposed which selects an optimum threshold. Method identifies three performance indices and establishes their relationship with threshold. A case study was done on the traffic data in San Antonio, Texas, which shows that the proposed method can

achieve a compression ratio of 8.12% of what the existing system provides. And the overall compression ratio is observed to be <1%. Analysis indicate that the selection of wavelet type does not affect the performance significantly, whereas higher decomposition levels result into better compression ratio.

Li et al, proposed an approach for flow volumes data compression for traffic network, based on Principal Component Analysis [21]. After preprocessing, PCA is applied on the dataset to break it into several principal components (PC). PCs are observed to have much less dimensions as compared to the original data. The data is recovered with the compression ratio of 6.2% and the recovery error of 13%.

Ding et al, proposed a an effective method for urban traffic data compression based on Wavelet Principal Component Analysis(PCA) [20]. In the proposed algorithm, the data is decomposed using wavelet and then multi-scale PCA is applied to reduce the dimension of the dataset. A simulation is performed for testing the algorithm and results show that this wavelet based PCA outperforms the conventional PCA algorithm for data compression.

Xiao-fa Shi proposed a real-time data compression and reconstruction method combining lifting wavelet transform and hybrid entropy coding [25]. In this approach, data is first decomposed with biorthogonal filters based on lifting wavelet transform and then the hard-threshold is applied to the wavelet coefficients. In order to achieve higher compression ratio, a hybrid entropy coding scheme is employed to code the reserved coefficients. This method combines both lossy and loss less compression techniques.

5.4 Introduction to EZW and SPIHT Algorithms

EZW (Embedded Zerotrees of Wavelet Transforms) and SPIHT (Set partitioning in hierarchical trees) are two well known lossy image compression algorithms. The EZW and SPIHT encoder is based on progressive encoding to compress an image into a bit stream with increasing accuracy. Hence when more bits are added to the stream, the decoded image contains more detail. At lower bit rates, most of the coefficients obtained after wavelet transform are zero or very close to zero. The encoder is based on two important observations:

- Generally images have a low pass spectrum. When an image is wavelet transformed the energy in the subbands decreases as the scale decreases. Hence the wavelet coefficients will be smaller generally, in the higher subbands than in the lower subbands. As the higher subbands only add detail, progressive encoding is a very good choice for compressing wavelet transformed images.

- Larger wavelet coefficients are more important than smaller ones.

EZW encoding scheme uses this two properties and codes the coefficients in decreasing order, in several passes. For each pass a threshold is chosen and coefficients are compared with it. If a wavelet coefficient is larger than the threshold, it is encoded and removed from the image, if it is smaller it is left for the next pass. When all the wavelet coefficients have been compared, the threshold is lowered and the image is scanned again to add more detail to the image. This process is repeated until all the wavelet coefficients have been encoded or another stopping criterion has been met.

5.5 1-Dimensional Data Compression

5.5.1 Data description

The flow detector data that we receive is updated at every five minutes. Flow detectors are installed at freeways. Data contains time stamp, traffic volume, average speed, detector ID and lane occupancy.

Data is extracted for a particular detector, for for a single day. For this research, only volume of vehicles is stored in the file. File looked like table 5.1.

Time(in minutes)	Volume(Veh/Hour)
2	144
7	84
12	180
17	84
22	132
27	108
32	108
37	96
42	72
47	72
52	48
..	..

Table 5.1: Data from Flow Detector

5.5.2 Framework and Methodology

How do we get compression in data using wavelet transform?

Irrespective of what wavelet or what upto what level we decompose, we can always construct the original signal using all the wavelet and scaling coefficients. Although number of coefficients (wavelet+scaling) is equal to the number of samples in input signal still we get a very good compression ratio because number of scaling coefficients, having large values, are $2K$ times less than original signal samples. Where K is the level of decomposition.

Whereas rest coefficients are wavelet coefficients, which represent the difference or the detail, are much lower in value, as compared to scaling coefficients. The distribution of values for the wavelet coefficients is generally centered around zero, with very few large coefficients. In other words, almost all the information is concentrated in a small fraction of the coefficients and can be efficiently compressed in lesser number of bits. This can be done by quantizing the values based on the histogram and encoding the result using an appropriate encoding technique e.g. Huffman Encoding or Embedded Zerotree Encoding. We can take a simpler route by discarding smaller coefficients and considering only M most significant coefficients. The compression ratio in this case will be $2M/N$. Factor of two arises due to the need of storing both coefficient index and its value.

Choosing the optimum level of decomposition and wavelet:

- The signal can be recursively decomposed to get finer detail and more general approximation. This is called multi-level decomposition. A signal can be de-

composed as many times as it can be divided in half. Thus, we only have one approximation signal at the end of the process.

- The low and high pass filters are basically the wavelet that is used. There have been a wide variety of wavelets created over time. The low pass is called the scaling function and the high pass is called the wavelet function. Different wavelets give different results depending on the type of data.

In order to find the best combination of wavelet type, decomposition level and threshold value, an exhaustive search is done trying many permutations and combinations (with the help of MATLAB)

5.5.3 Performance Measures and Parameters

Threshold Selection

There are two types of thresholding:

- Hard thresholding (shrink or kill)

$$Thr = \begin{cases} median(abs(detail\ at\ level\ 1)); & if\ nonzero \\ 0.05\ max(abs(detail\ at\ level\ 1)); & otherwise \end{cases} \tag{5.1}$$

  In hard thresholding, the coefficients below a certain threshold are set to zero and the magnitudes of the wavelet coefficients above the threshold are left un-

changed.

$$T_d^{hard} = \begin{cases} d\ ; & |d| > Thr \\ 0\ ; & |d| \leq Thr \end{cases}$$

- Soft thresholding (keep or kill)

$$Thr = \sqrt{2.log(n)} \tag{5.2}$$

$$\text{and} \quad n = prod(size(x))$$

In soft thresholding, the coefficients below a certain threshold are set to zero whereas the remaining coefficients are reduced by an amount equal to the value of the threshold.

$$T_d^{soft} = \begin{cases} sgn(d)(|d| - Thr)\ ; & |d| > Thr \\ 0\ ; & |d| \leq Thr \end{cases}$$

Hard thresholding is default for compression whereas soft thresholding is recommended for denoising of a given signal. For analysis purpose, we have used both, hard and soft thresholding, and compared the results.

Percentage Recovery

It is the ratio of the L^2 norm squared of the thresholded wavelet coefficients reconstructed signal to the L^2 norm squared of the original wavelet coefficients (which is equal to the signal L^2 norm squared) expressed as a percentage. If this number is

close to 100, it means that the energy in denoised or compressed version is very close to the energy in the uncompressed version. If X is a one-dimensional signal then percentage recovery is reduced to -

$$\text{Percentage Recovery}(R) = \frac{\|XC\|^2}{\|X\|^2} 100$$

Compression Score

It is the ratio of the number of thresholded wavelet coefficients that are zero to the total number of wavelet coefficients, expressed as a percentage. Compression Score near 100 here means that virtually all our thresholded wavelet coefficients are zero, meaning that our signal or image is being reconstructed based on a very sparse set of wavelet coefficients.

5.5.4 Results and Discussion

Table 5.2 and 5.3 show the compression ratio and percentage recovery of the data for level independent(hard) and level dependent(soft) thresholding respectively. We observe that in case of level independent thresholding percentage recovery is above 99% for all the levels and wavelets, while the compression ratio ranges from 25.35% to 54.52%. Whereas in case of level dependent thresholding compression ratio ranges from 25.00% to 49.48%

Level Independent(Hard) Thresholding				
Wavelet	Level	Recovery (%)	Compression Ratio (%)	Threshold (absolute value)
db3	1	99.84	27.05	25.45
	2	99.83	26.35	25.45
	3	99.83	27.00	25.45
	4	99.89	25.35	25.45
	5	99.82	29.31	25.45
	6	99.76	42.57	25.45
db5	1	99.76	41.78	25.45
	2	99.75	41.99	25.45
	3	99.81	40.28	25.45
	4	99.77	43.69	25.45
	5	99.73	48.84	25.45
	6	99.73	47.76	25.45
db7	1	99.72	49.85	25.45
	2	99.78	46.53	25.45
	3	99.73	49.49	25.45
	4	99.72	53.11	25.45
	5	99.71	51.09	25.45
	6	99.71	51.78	25.45
haar	1	99.77	48.96	25.45
	2	99.72	52.86	25.45
	3	99.73	54.05	25.45
	4	99.74	52.28	25.45
	5	99.73	54.00	25.45
	6	99.76	49.31	25.45
sym2	1	99.72	54.52	25.45
	2	99.77	53.50	25.45
	3	99.79	52.07	25.45
	4	99.79	53.99	25.45
	5	99.77	49.48	25.45
	6	99.74	54.30	25.45

Table 5.2: Results-Level Independent Thresholding

Level Dependent(Soft) Thresholding				
Wavelet	Level	Recovery (%)	Compression Ratio (%)	Threshold (absolute value)
db3	1	99.87	25.00	23.47
	2	99.85	25.00	24.67
	3	99.86	25.00	23.19
	4	99.89	25.35	25.46
	5	99.89	25.17	20.80
	6	99.84	37.50	21.81
db5	1	99.84	37.17	23.17
	2	99.83	37.18	21.60
	3	99.85	37.50	24.00
	4	99.86	37.54	20.74
	5	99.81	43.52	21.82
	6	99.81	43.27	23.06
db7	1	99.82	42.77	20.48
	2	99.83	43.75	24.00
	3	99.83	43.73	21.21
	4	99.82	46.56	21.62
	5	99.79	46.11	23.27
	6	99.81	45.56	21.48
haar	1	99.79	47.57	24.00
	2	99.82	46.80	21.25
	3	99.82	47.90	22.00
	4	99.81	47.42	23.45
	5	99.82	47.14	21.36
	6	99.78	48.61	24.00
sym2	1	99.82	48.16	22.11
	2	99.83	48.73	22.79
	3	99.83	48.22	23.50
	4	99.85	47.66	21.61
	5	99.77	49.48	25.46
	6	99.82	49.01	22.65

Table 5.3: Results-Level Dependent Thresholding

Figure 5.1, 5.2, 5.3 and 5.4 show the original signal, reconstructed signal, approximation and wavelet coefficients for the best results based on tables 5.2 and 5.3.

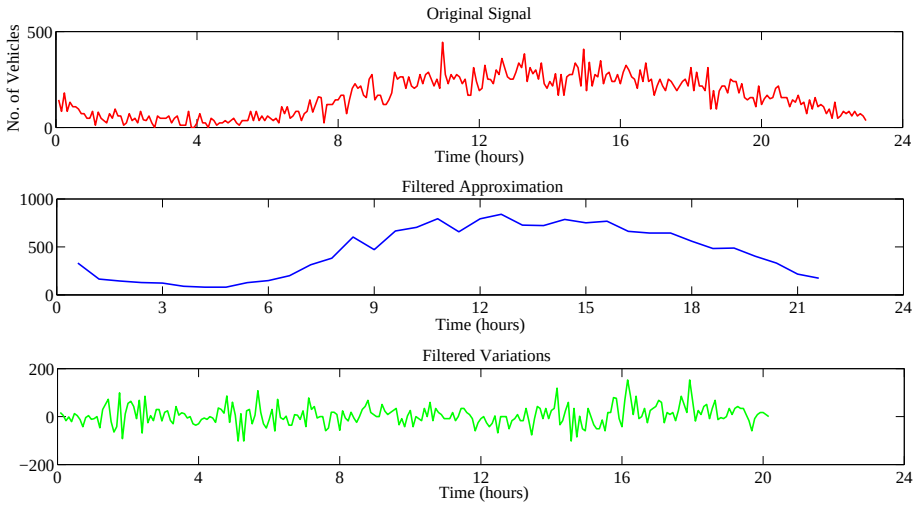

Figure 5.1: Haar-Level3

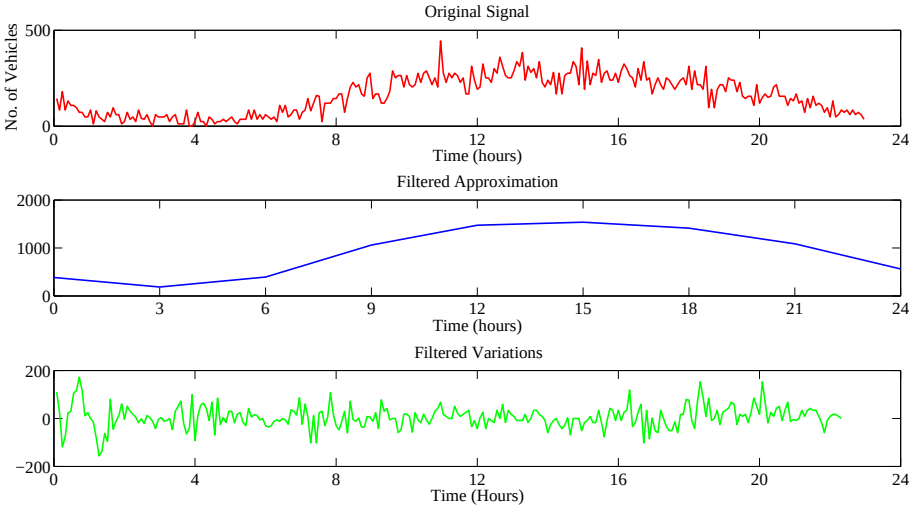

Figure 5.2: Haar-Level5

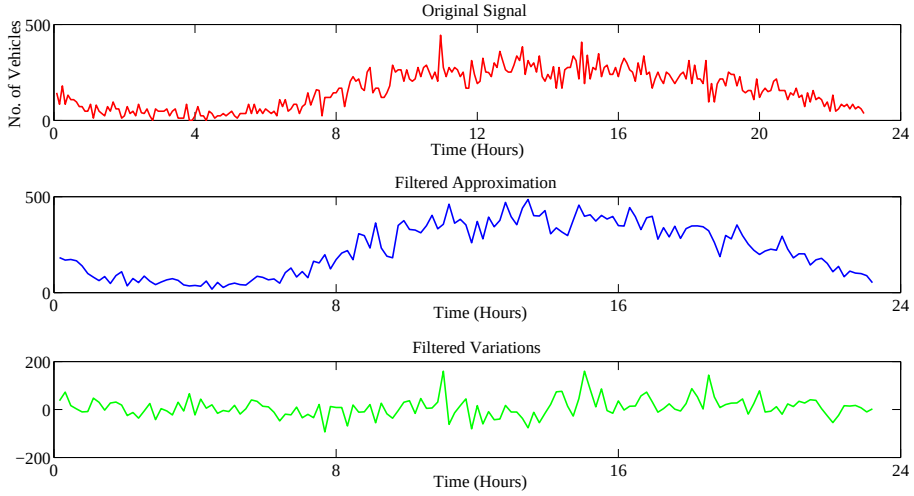

Figure 5.3: Svm2-Level1

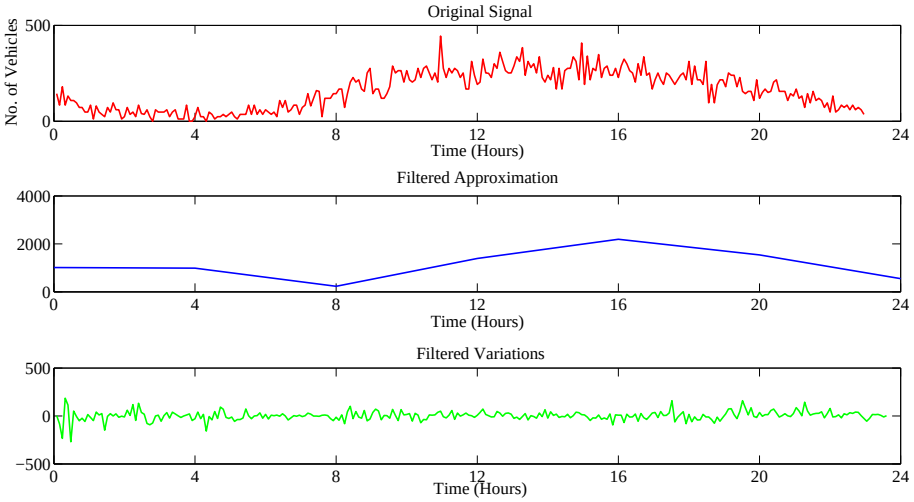

Figure 5.4: Sym2-Level6

5.6 2-Dimensional Data Compression

5.6.1 Proposed Approach

We rearrange our data in a 2 dimensional matrix. It is formed by combining several days data into one file. Table 5.4 shows one such arrangement where we arrange volume data in a 2D matrix. This matrix can be treated as a gray scale image. We convert this matrix into a .bmp file and apply image compression techniques to compress the traffic data.

Days	Time								
Day 1	144	84	180	84	132	108	108	96	...
Day 2	72	24	84	36	120	60	48	12	...
Day 3	36	12	96	48	12	48	72	48	...
Day 4	24	36	72	36	12	12	12	60	...
Day 5	96	12	60	60	48	12	36	36	...
Day 6	48	96	60	72	60	48	24	60	...
Day 7	90	60	48	108	60	120	60	60	...
...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...

Table 5.4: 2-D Data Arrangement

5.6.2 Data Visualization

With the help of color coding we can visualize this 2-D data effectively. Figure 5.5 shows a color coded figure representing volume during a day, spanning 16 days.

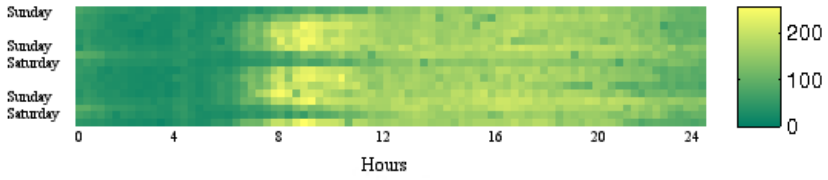

Figure 5.5: 2D-visualization

5.6.3 Performance Measures and Parameters

Reduction Ratio

A measure of achieved compression is given by the compression ratio (CR) and the Bit-Per-Pixel (BPP) ratio. CR and BPP represent equivalent information. CR indicates that the compressed image is stored using CR % of the initial storage size while BPP is the number of bits used to store one pixel of the image. Reduction Ratio indicates the reduction in percentage, of the initial storage size.

$$CR(\%) = \frac{\text{Size of Compressed Image}}{\text{Size of Original Image}} \cdot 100 \tag{5.3}$$

$$RR(\%) = 100 - CR(\%)$$

Error Metrics

Two of the error metrics used to compare the various image compression techniques are the Mean Square Error (MSE) and the Peak Signal to Noise Ratio (PSNR). The MSE is the cumulative squared error between the compressed and the original image, whereas PSNR is a measure of the peak error.

5.6.4 Data Compression using EZW and SPIHT

We make a 2D data matrix consisting of 16 days data represented by 16 columns and time during the day on the rows as shown in table 5.4. Basically we have a 64X16 data matrix, which we treat as a gray scale image for compression. Figure 5.6, 5.7 and 5.8 show results obtained by using EZW and SPIHT algorithm on a 2D data matrix for compression.

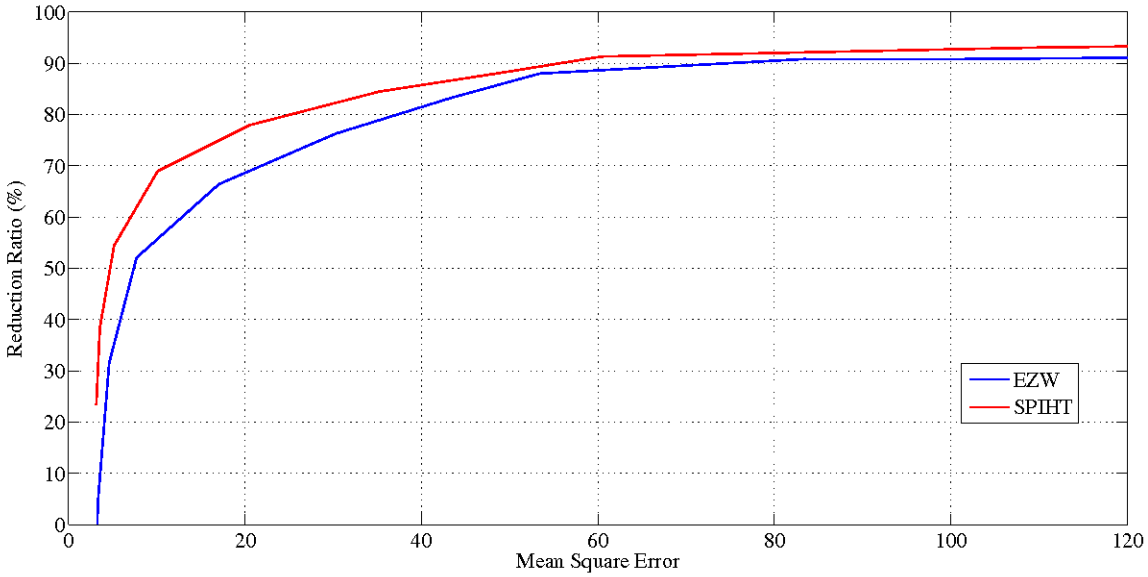

Figure 5.6: 2D-Compression of Occupancy Data

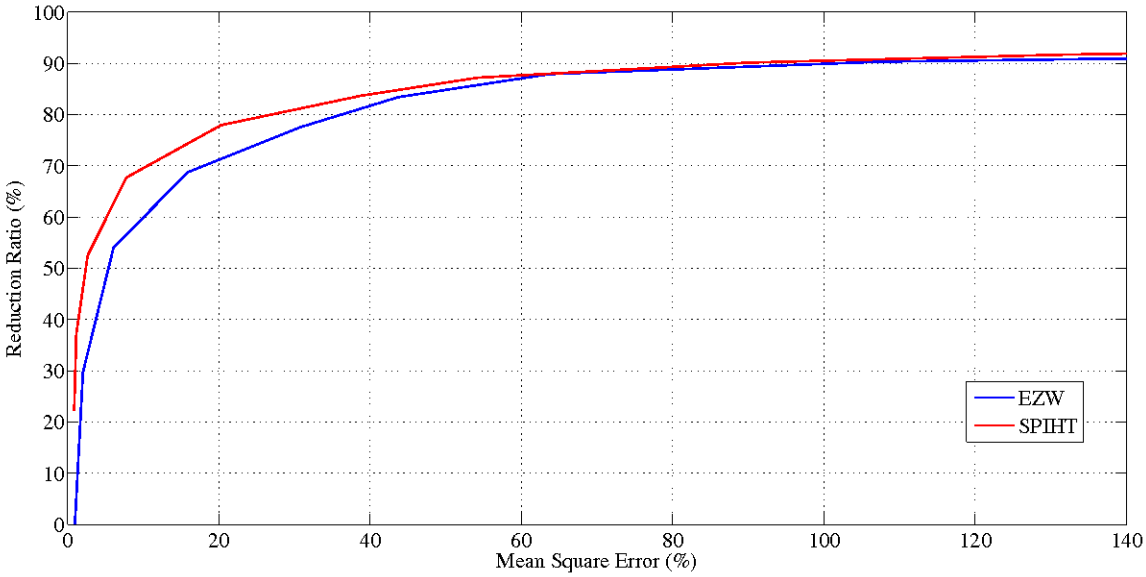

Figure 5.7: 2D-Compression of Speed Data

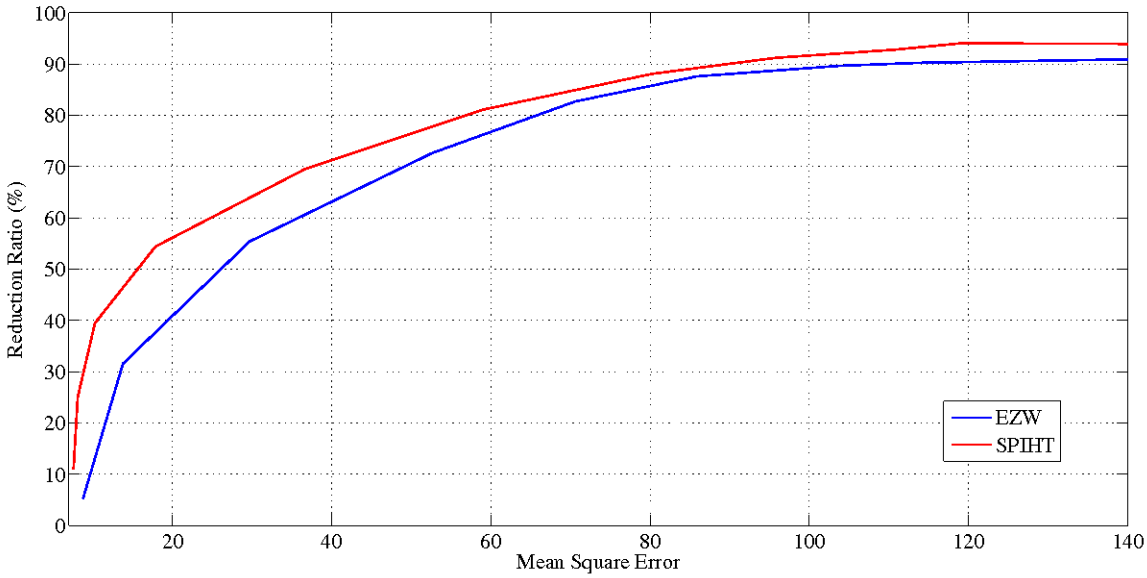

Figure 5.8: 2D-Compression of Volume Data

We observe that for all three traffic parameters: volume, speed and occupancy, SPIHT encoding outperforms EZW. Figure 5.9 compares performance of SPIHT algorithm for different type of traffic data.

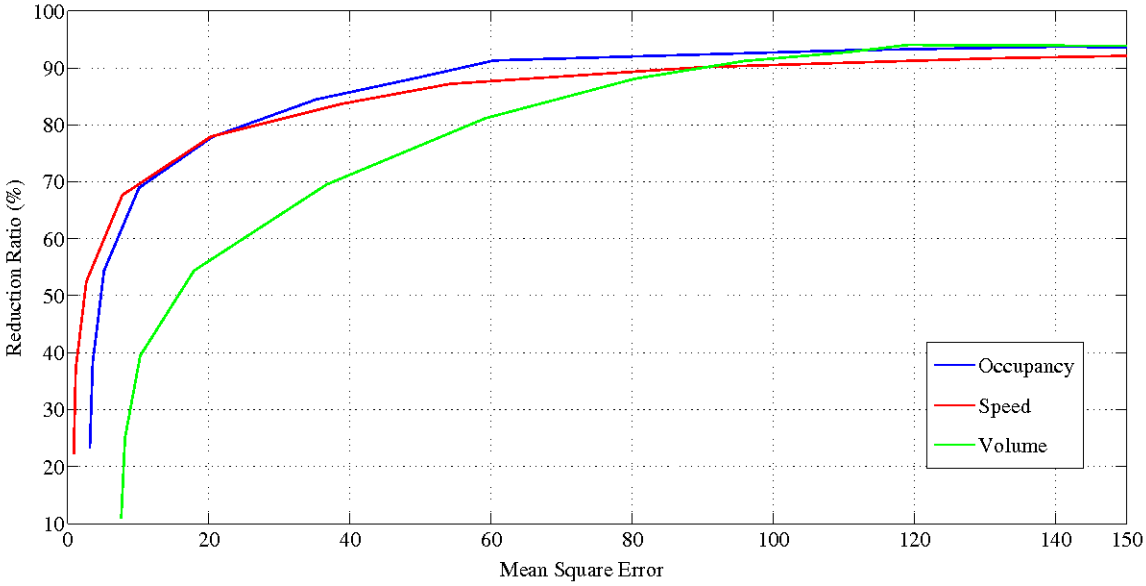

Figure 5.9: Comparison using SPIHT

5.7 3-Dimensional Data Compression

5.7.1 Proposed Approach

In the previous section, we proposed a data compression technique for traffic data, where we rearrange the data in a 2D matrix and treat them as a gray scale image to perform EZW and SPIHT algorithms. In this section we propose a compression technique, which uses a 3-Dimensional matrix, made by combining three 2-D matrix. We take three 2-D matrix, containing volume, speed and occupancy data as described in previous section, and combine them as a 3D matrix. Now it is basically 64X16X3 matrix, and we treat it as a colored image and apply SPIHT compression techniques on it. Figure 5.10 shows the result for the compression.

5.7.2 Results

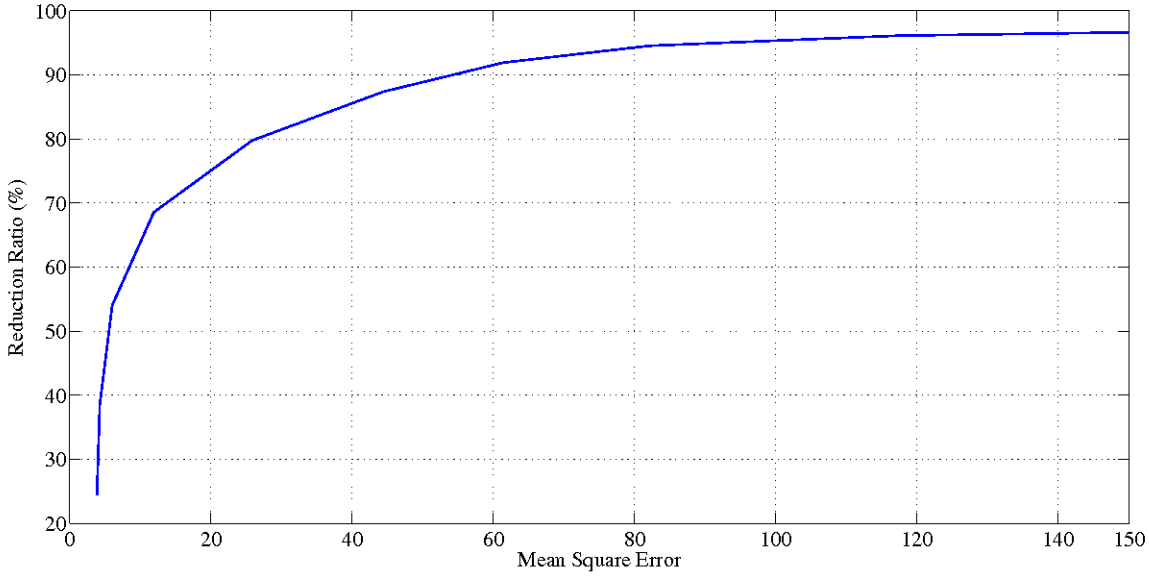

Figure 5.10: 3-D Data Compression using SPIHT

Figure 5.11 compares of 2D data matrices with the combined 3D data matrix.

Figure 5.12 compares compression of 2D vs 3D Volume data arrangement.

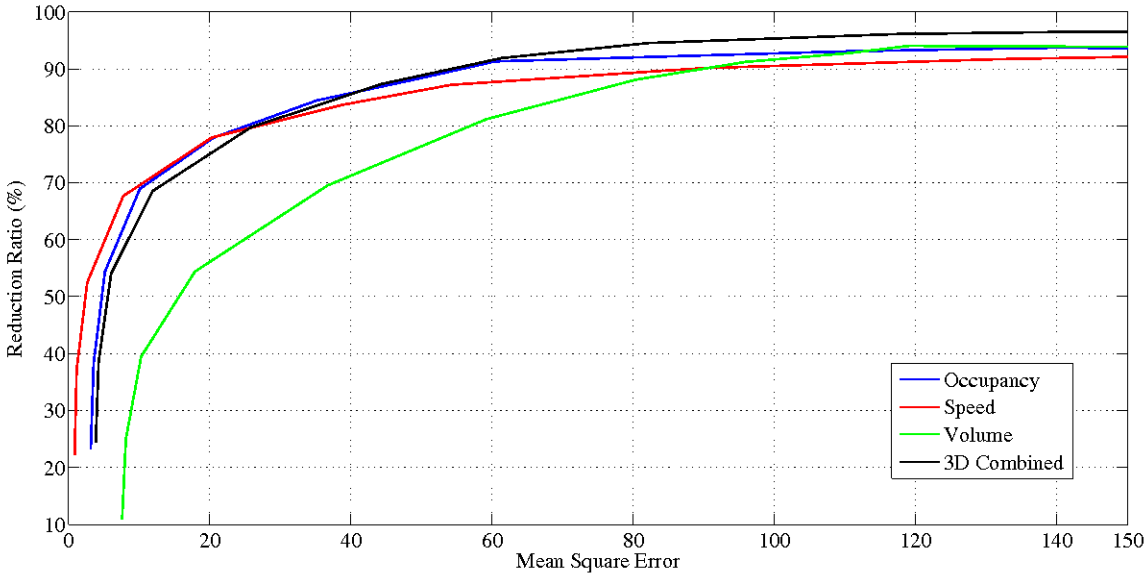

Figure 5.11: Comparison of 2D and 3D compression using SPIHT

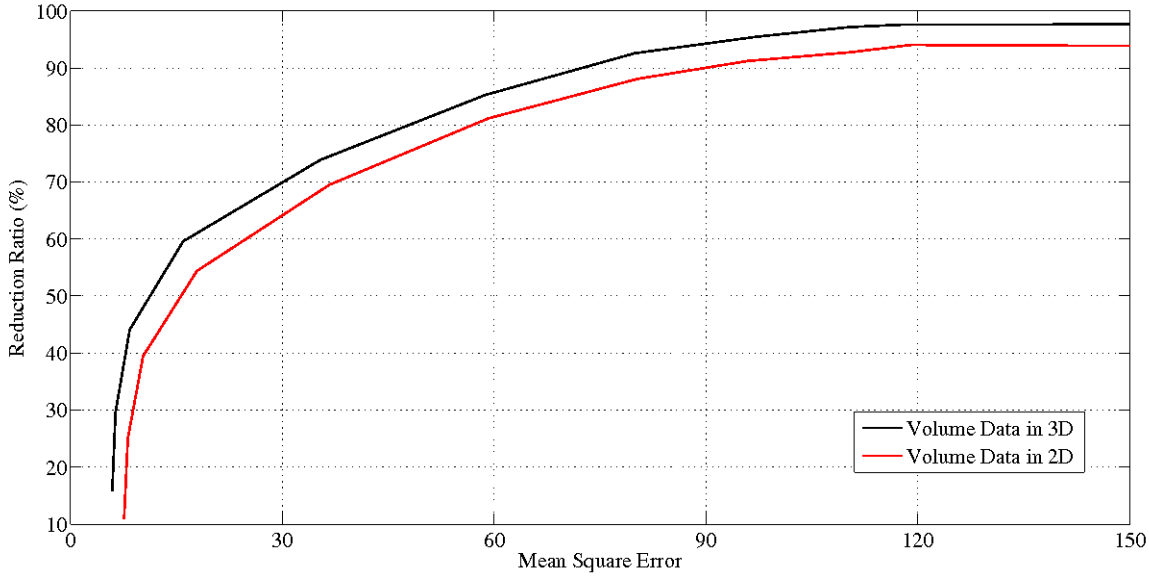

Figure 5.12: Compression of 2D vs 3D Volume Data using SPIHT

5.8 Conclusion

This chapter focused on use of wavelet transform in transportation data compression. We investigated compression of loop detector data. We divided our analysis into three sections based on the data arrangement. First data arrangement, named 1D, was basically the most conventional data arrangement used for storing the loop detector data. The file containing the data had all the traffic parameters for a single detector, varying over the time. We used wavelet transform on this data for visualization of overall daily traffic pattern and its compression. We investigated various wavelets and used various level of decomposition to analyze their performance.

Second data arrangement, named 2D, was obtained by rearranging and combining several days data into a single 2D matrix. We treated this matrix as a grayscale image and investigated EZW and SPIHT compression techniques for its compression. We observed that for all the traffic parameters, SPIHT outperformed EZW encoding. While occupancy and speed had almost overlapping compression vs mean square error curve, volume was least compressible among all.

Third data arrangement, named 3D, was obtained by combining three 2D matrices, and treating them as a colored image. We repeated the investigation for 3D arrangement same like we did for 2D arrangement. We observed that the compression performance for 3D was slightly better than for the 2D matrix arrangement.

CHAPTER 6

LOGISTIC REGRESSION-WAVELET MODEL FOR TRAFFIC INCIDENT DETECTION

6.1 Introduction

Traffic incidents are non periodic and pseudo random events causing traffic jam and hitting the overall performance of the road network. Probability of traffic incidents is higher during the peak hours. Many major cities in US have a traffic management system which includes traffic characteristic detectors and a centralized operations center for monitoring. These detectors comprise of video cameras, bluetooth sensors, flow detector sensors etc. They can capture traffic characteristics, such as traffic speed, occupancy and volume. However automatic incident detection techniques using these data are not widely used yet. Reliable and quick detection of incidents can prove very useful in incident management on roadways. Emergency crew can be sent on the incident location for obstruction clearance and medical help. It will also help to manage detour efficiently and better management of traffic and road network in case of an incident.

Many researchers have worked on the problem of real time incident detection techniques using the real time traffic data. However the main issue is the confidence level with which the incident is predicted. It might happen when a incident happens

but the algorithm is unable to detect it or vice versa (false alarm). Incident detection can be seen as a classification problem with two outcomes: Incident detected or not detected. Based on the available data (volume, occupancy, speed etc), algorithms decide whether the data represents an incident or not. Misclassification of either of the two reduces the reliability and usability of the system. So the objective is to develop a reliable automatic traffic incident detection system which analyzes the data and predicts the incidents efficiently and with high confidence.

In this chapter we will focus on regression model for dichotomous data, i.e. logistic regression. This model is suitable when the outcome can takes only limited number of values, in our case only two, presence or absence of an incident. We will look into generalized linear model (gle) with binomial response and logit link function. We will present a framework to use logistic regression for incident prediction in transportation systems. Test the model on the historical traffic data and analyze the reliability and robustness of the system.

Further in the chapter we investigate effect of denoising of the data using wavelets, on the performance of incident detection model. This will also help us, in assessing the effects of traffic data compression using wavelets. Literature suggests that for some of the incident detection algorithms, feature extraction using DWT actually increases the detection rate. We will study in detail effect of DWT on logistic regression models for traffic incident detection.

6.2 Literature Survey

Many algorithms have been proposed for traffic incident detection over the years. The information, that we get generally from traffic sensors is about traffic occupancy, speed, volume and flow rate etc. Traffic occupancy indicates the fraction of time a particular location is occupied by a vehicle and flow rate indicates the number of vehicles passing through a location in unit time. The methods proposed for incident detection range from simple threshold comparisons to more complex model based predictions.

Recently researchers shifted their focus towards model-free detection techniques involving fuzzy logic theory, neural networks or a combination of both. In the fuzzy logic approach, objective is to build a fuzzy knowledge with the available historical data and come up with some fuzzy rules. These rules are then processed by fuzzy logic system to identify and predict the outcomes. Chang and Wang evaluated the applications of fuzzy set theory to improve existing incident detection algorithms [13]. They compared system performance with that of conventional systems using real-world volume and occupancy data, and summarized benefits and needed improvements in the existing incident detection algorithms to utilize fuzzy logic theory.

Artificial Neural Networks (ANNs) are known to be powerful in pattern recognition and classification problems. They act like a model-free black box. They are adaptive and grab the structure of data quickly and efficiently. Cheu and Ritchie proposed a methodology for automated detection of lane-blocking freeway incidents using artificial neural networks (ANNs) [14]. They developed three types of neural

network models, namely the multi-layer feedforward (MLF), the self-organizing feature map (SOFM) and adaptive resonance theory 2 (ART2), to classify the traffic data. MLF was found to give best results, among the three ANNs, to construct a better incident detection system.

The conjugate gradient method uses the steepest descent technique, where weight changes are made along the direction, resulting in minimization of system error. Adaptive Conjugate Gradient Neural Network Model (ACGNN) involves adaptive conjugate gradient algorithm for training neural networks. Adeli and Samant investigated ACGNN model for traffic incident detection problems [15]. They tried the algorithm with various combination of traffic data series such as traffic volume, speed and occupancy. Results showed the best incident detection rate of 91.1% with combination of all three and the false alarm rate of 5.1%. Further enhancement was done by combining DWT and Linear Discriminant Analysis(LDA) with ACGNN [12]. The new computational model was based on preprocessing the traffic data by DWT and LDA, followed by ACGNN. Results were much better with a high detection rate of 97.8% and a lower false alarm rate of 1%.

Samant and Adeli investigated methodology of enhancing fuzzy neural network based traffic incident detection algorithms using wavelets [16]. They showed that the performance of a fuzzy neural network algorithm can be improved through preprocessing of data using a wavelet-based feature-extraction model. In this approach discrete wavelet transform (DWT) denoising and feature-extraction model was combined with the fuzzy neural network approach. It was observed that use of the wavelets for de-

noising the traffic data helps in increasing the incident-detection rate, reducing the false-alarm rate and the incident-detection time, and improving the convergence of the neural network training algorithm to a large extent.

6.3 Logistic Regression

Logistic Regression is type of regression analysis where we can predict categorical outcomes based on certain predictors. Probabilities of the possible outcomes are modeled, using logistic functions, as a function of independent variables. Logistic regression can be binomial or multinomial. Logistic regression uses a link function which transforms the limited range of a probability $[0, 1]$, into $(-\infty, +\infty)$.

6.3.1 Bernoulli and Binomial Distributions

Lets consider a response y_i which is binary and has the following value:

$$y_i = \begin{cases} 1, & \text{if there is a traffic incident} \\ 0, & \text{otherwise} \end{cases}$$

y_i can be seen as the realization of the random variable Y_i with two outcomes one or zero with probabilities p_i and $1 - p_i$ respectively. With this arrangement the distribution of Y_i is called Bernoulli distribution [17]. It can be represented as:

$$Pr\{Y_i = y_i\} = p_i^{y_i} (1 - p_i)^{1-y_i} \tag{6.1}$$

Expected value and variance of this distribution are given by:

$$E(Y_i) = \mu_i = p_i \quad \text{and} \quad var(Y_i) = \sigma_i^2 = p_i(1 - p_i) \tag{6.2}$$

We observe that variance is dependent on the probability p_i. This means any factor affecting p_i will also change the variance. Hence linear models are are not suitable for this case as they assume variance to be independent of predictors.

Now lets assume that our data is divided into k groups, with n_i denoting the number of observations on i^{th} group and y_i denoting the number of favorable outcomes. We can again see y_i as the realization of random variable Y_i which can take values 0,1,2,....n_i. If n_i observations in each group are independent having same probability p_i, the this distribution is called Binomial distribution and represented as:

$$Y_i = B(n_i p_i).$$

and its probability distribution function is given as:

$$Pr\left\{Y_i = y_i\right\} = \binom{n_i}{y_i} p_i^{y_i} \left(1 - p_i\right)^{1 - y_i} \tag{6.3}$$

Expected value and variance of this distribution are given by:

$$E(Y_i) = \mu_i = n_i p_i \quad \text{and} \quad var(Y_i) = \sigma_i^2 = n_i p_i(1 - p_i) \tag{6.4}$$

In this case also the variance is dependent on probability p_i and in return on the predictors affecting probability. Hence linear models are are not suitable for this case also.

6.3.2 Logit Transformation

Now we will try to build a model for our data. We want the probabilities (p_i) to depend on the predictors or covariates (x_i). To start with we can simple assume a linear model where (p_i) is a linear function of (x_i) as:

$$p_i = \beta x_i' \tag{6.5}$$

where β is a vector of regression coefficients. This model is called linear probability model and it is generally estimated using ordinary least squares methods (OLS). Problem with this model is that probability on the left hand side can take value only between zero and one, whereas the right hand side can take any real value. We solve this problem by following two steps. Instead of probability, we consider odds as:

$$Odds_i = \frac{p_i}{1 - p_i}$$

and next we take log of this odd to get the logit or log-odds:

$$\eta_i = logit(p_i) = log\frac{p_i}{1 - p_i} \tag{6.6}$$

η_i has the range from $+\infty$ to $-\infty$.

6.3.3 Logistic Regression Model

Suppose the independent observations y_i are realization of random variable Y_i having a binomial distribution

$$Y_i = B(n_i p_i) \tag{6.7}$$

Now if we assume that logit of the probability p_i is dependent linearly on the predictors then we can write

$$logit(p_i) = \beta x_i' \tag{6.8}$$

Model defined by the equations 6.7 and 6.8 form a generalized linear model (gle) with binomial response and link logit.

6.3.4 Model Fitting and Hypothesis testing

Maximum Likelihood Estimation

Logistic regression uses the maximum likelihood procedure for estimating the coefficients. Based on the conditions and predictors this procedure maximizes the likelihood of the regression coefficients. Maximum likelihood estimation is an iterative process which starts with an initial guess and repeats the process until the model converges. In this section we give a brief outline of procedure for calculating the best estimate of the vector of regression coefficients β. The likelihood function for n independent binomial observations is given by equation 6.3. From this equation we compute the

log-likelihood function, which comes out to be

$$logL(\beta) = \sum \{y_i log(p_i) + (n_i - y_i) log(1 - p_i)\}$$

We maximize the log likelihood function by iteratively re weighted least squares (IRLS). If we have an estimate $\hat{\beta}$, we calculate linear prediction from it

$$\hat{\eta} = x_i{}'\hat{\beta}$$

and the fitted values

$$\hat{\mu} = logit^{-1}(\eta)$$

From these estimated coefficients we define a working dependent variable z as

$$z_i = \eta_i + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i(n_i - \hat{\mu}_i)}$$

Weighted least square estimate is is now calculated by regressing z on the covariates

$$\hat{\beta} = (X'WX)^{-1}X'Wz$$

where W is a diagonal matrix of weights having elements

$$w_{ii} = \frac{\hat{\mu}_i(n_i - \hat{\mu}_i)}{n_i}$$

The resulting estimate of β is used to obtain best fit values and iteration is done till it converges.

Deviance and Likelihood Ratio Tests

In logistic regression analysis, deviance is analogous to the sum of squares calculations in linear regression. It gives a measure of the lack of fit of the data in the logistic regression model. Deviance is calculated by the likelihood ratio test, comparison of a given model with a model with a theoretically perfect fit (saturated model):

$$Deviance(D) = -2\ ln\ \frac{\text{likelihood of the fitted model}}{\text{likelihood of the saturated model}}$$

Smaller values of deviance indicates lesser deviation from the saturated model and hence a better fit.

Hypothesis Testing

After estimating goodness of fit of the model we want to know significance of each of the predictors on the outcome. In linear regression, the significance of a regression coefficient is assessed by t-test. In logistic regression, the tests for assessing the significance of an individual predictor most common tests are the likelihood ratio test and the Wald statistic which are analogous to computing t-test in linear regression. We have already discussed likelihood ratio test in above section. Wald statistic is computed by the ratio of the square of the regression coefficient to the square of the

standard error of the coefficient:

$$W_j = \frac{\beta_j^2}{SE_{\beta_j}^2}$$

where SE denotes the standard error of the coefficients.

Consider a model matrix and vector of coefficients

$$X = (X_1, X_2, ...X_j, ...) \;\; and \;\; \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ ... \\ \beta_j \\ ... \end{pmatrix}$$

now consider testing the hypothesis that variables in X_2 have no effect on the response

$$H_0 : \beta_j = 0$$

by calculating the ratio of estimate to its standard error

$$z = \frac{\hat{\beta}_j}{\sqrt{\hat{var}(\hat{\beta}_j)}}$$

We can use Wald test to determine the confidence interval of β_j. We can say with

a confidence of $100(1-\alpha)\%$ that β_j lies in the interval

$$\hat{\beta}_j \pm z_{1-\alpha/2}\sqrt{\hat{var}(\hat{\beta}_j)}$$

where $z_{1-\alpha/2}$ is normal critical value for a two sided test of size α.

6.4 Data Source and Characteristics

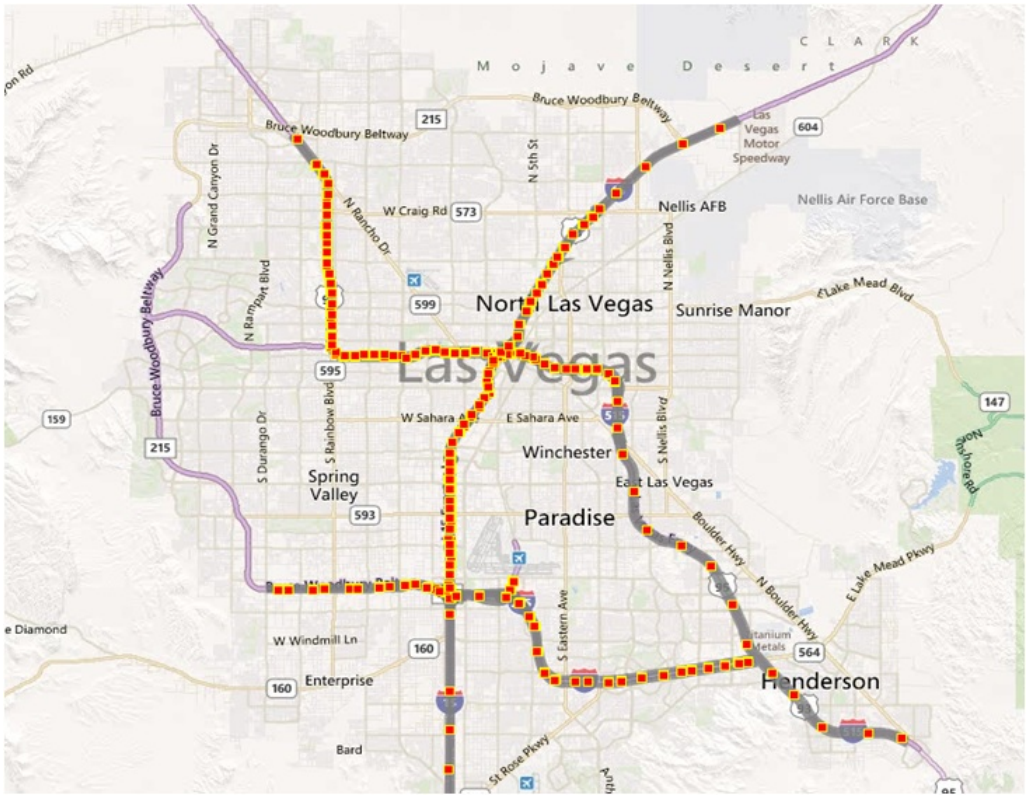

Figure 6.1: Traffic Sensors on freeway (Las Vegas Area)

Figure 6.1 shows the traffic sensors placed along the freeways US-95 and I-15 in the

Las Vegas area. This data can be downloaded from FAST website <http://rtcsnv.com /mpo/fast/dashboard.cfm>. This data contains lane wise speed, occupancy, volume along with the time stamp. This data is averaged and updated at every fifteen minutes interval. This data is available for each of the detectors placed along the freeway and can be downloaded separately for each of them.

Another database on FAST's website gives details about the time and location of the traffic incidents on the freeway. We analyzed the data and identified the location with maximum number of incidents as I-15 North bound, past Sahara (fig 6.2). We downloaded two separate files containing traffic parameters and traffic incidents during April 2012. We then combined the two data sets by matching the time stamps and formed one single database that looked like table 6.1.

Occupancy	Volume	Avg Speed	Incident
5	1848	61.6	0
5.8	1607	60.2	0
5.2	1840	63	0
5.2	1805	63.6	0
3.8	1945	36	1
4	1652	21.2	1
3.6	1744	23.8	1
3	1649	21	1
2.6	1770	22.4	1
3.2	1770	25.6	0
2.8	1866	24.6	0
2.4	1206	25.8	0
2.4	1474	28.6	0
2.4	1845	25.8	0
2	1971	27.8	0
2.4	1825	24.8	0
2.4	1858	24.6	0
2.6	1794	28.2	0
2.2	1919	26.2	0
1.6	1768	45.8	0
5	1959	49.8	0
5.8	1727	60	0
5.2	1692	65.6	0

Table 6.1: Dataset for Incident Detection

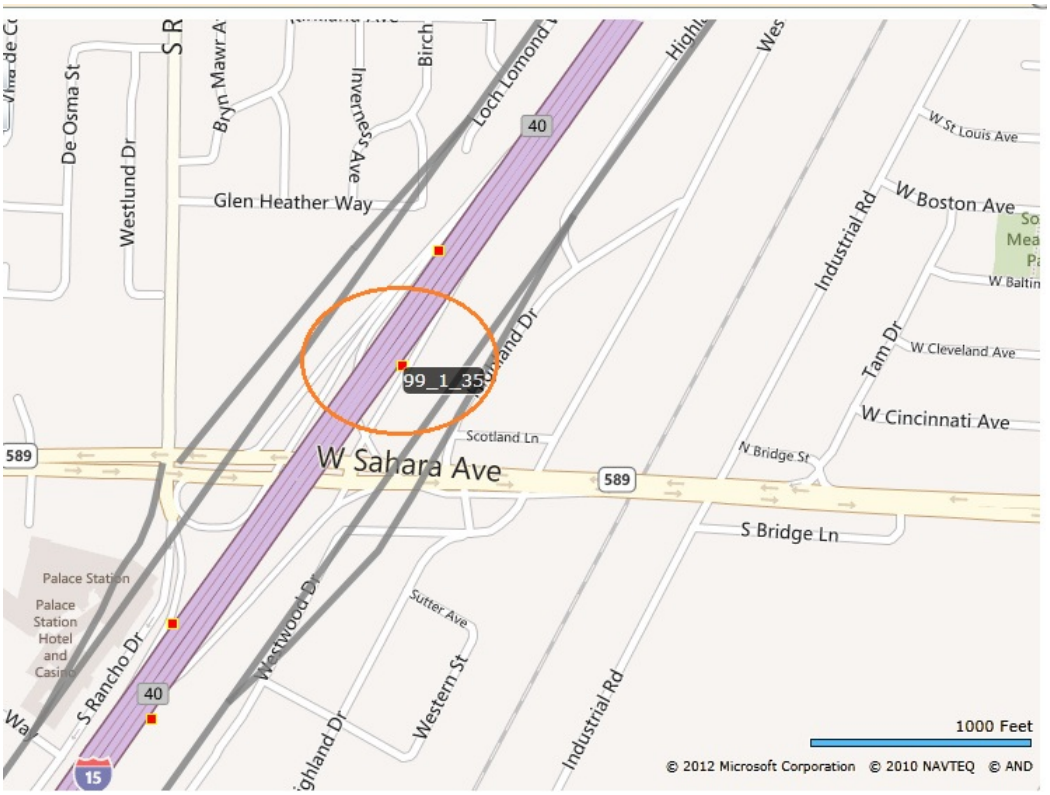

Figure 6.2: Identified Crash Site

6.5 Incident Detection Using Logistic Regression

6.5.1 Framework and Methodology

In the sample data set shown in table 6.1, incident column has the value either zero or one. Zeros indicate non incident or normal traffic behavior and ones indicate a traffic incident. To incorporate the prolonged effects of incidents like congestion, lane blocking etc even after the incident, we made incident column as one along with the actual sample in which the incident happened.

For our analysis we take help of the NLOGIT, which is integrated package for estimating and analyzing discreet choice models.

For estimating the model we use the following commands in NLOGIT:

LOGIT

;LHS = Dependent Variables

;RHS = Predictors $

Along with the traffic parameters we use a constant K, as the dependent variable.

We estimate the following parameters:

- $logL$ = log likelihood function at the maximum
- $logL_0$ = the log likelihood function assuming all slopes are zero
- Chi square statistic for testing hypothesis $H_0 : \beta$ and the significance level (probability that χ^2 exceeds test value). The statistic is:

$$\chi^2 = 2(logL - logL_0)$$

- Other results include coefficients estimates, standard errors, t-ratios etc.

6.5.2 Results and Discussion

Following are the results of logistic regression model applied on traffic data for incident detection using Avg Speed, Volume and Occupancy as predictors. Table 6.2 gives maximum likelihood estimation for the model.

Parameter / Test	Value
Dependent variable	Incident
Weighting variable	None
Number of observations	99
Iterations completed	8
Log likelihood function	-20.13
Restricted log likelihood	-42.10
Chi squared	43.95
Degrees of freedom	3
Prob[ChiSqd > value]	.00
Hosmer-Lemeshow chi-squared	2.83
P-value	.94 (8 df)

Table 6.2: Maximum Likelihood Estimates

Table 6.3 gives estimated values of coefficients and standard error for the variables.

Variable	Coefficient	Std Error	b/St.Er.	P[$\|Z\|$ >z]	Mean of X
Constant	.7329	4.0302	.182	.8557	
Occupancy	-.0673	.0966	-.697	.4860	11.07
Volume	.0034	.0014	2.384	.0171	1224.24
Speed	-.1392	.0610	-2.281	.0225	56.46

Table 6.3: Estimation of Coefficients and Std. Error

Table 6.4 gives frequencies of actual and predicted outcomes. Threshold value for predicting $Y = 1$ (incident) is 0.5.

	Predicted		
Actual	0	1	Total
0	78	6	84
1	4	11	15
Total	82	17	99

Table 6.4: Frequencies of actual and predicted outcomes

Analysis of Binary Choice Model predictions are based on threshold = 0.5. Table 6.5 and 6.6 give prediction success and prediction failure of the model.

Parameter	Description	Value (%)
Sensitivity	actual 1s correctly predicted	73.33
Specificity	actual 0s correctly predicted	92.85
Positive predictive value	predicted 1s that were actual 1s	64.70
Negative predictive value	predicted 0s that were actual 0s	95.12
Correct prediction	actual 1s and 0s correctly predicted	89.89

Table 6.5: Prediction Success

Parameter	Description	Value (%)
False pos. for true neg.	actual 0s predicted as 1s	7.14
False neg. for true pos.	actual 1s predicted as 0s	26.66
False pos. for predicted pos.	predicted 1s actual 0s	35.29
False neg. for predicted neg.	predicted 0s actual 1s	4.87
False predictions	actual 1s and 0s incorrectly predicted	10.10

Table 6.6: Prediction Failure

Table 6.7 gives the incident detection rate and false alarm rate for different of traffic parameters (using threshold probability as 0.5). We observe highest detection rate of 64.15% while using a combination of Volume + Speed + Occupancy.

Traffic Data	Incident Detection Rate (%)	False Alarm Rate (%)
Volume (Veh/h)	18.87	0.99
Occupancy (%)	50.94	5.59
Avg Speed (miles/h)	58.49	5.26
Avg Speed + Occupancy	60.38	4.61
Avg Speed + Volume	64.15	5.26
Occupancy + Volume	54.72	4.61
Avg Speed + Occupancy + Volume	64.15	5.26

Table 6.7: Incident Detection Results using Logit Models

Figure 6.3 shows the variation of incident detection rate against False alarm rate, as we vary the threshold probability from 1 to 0.

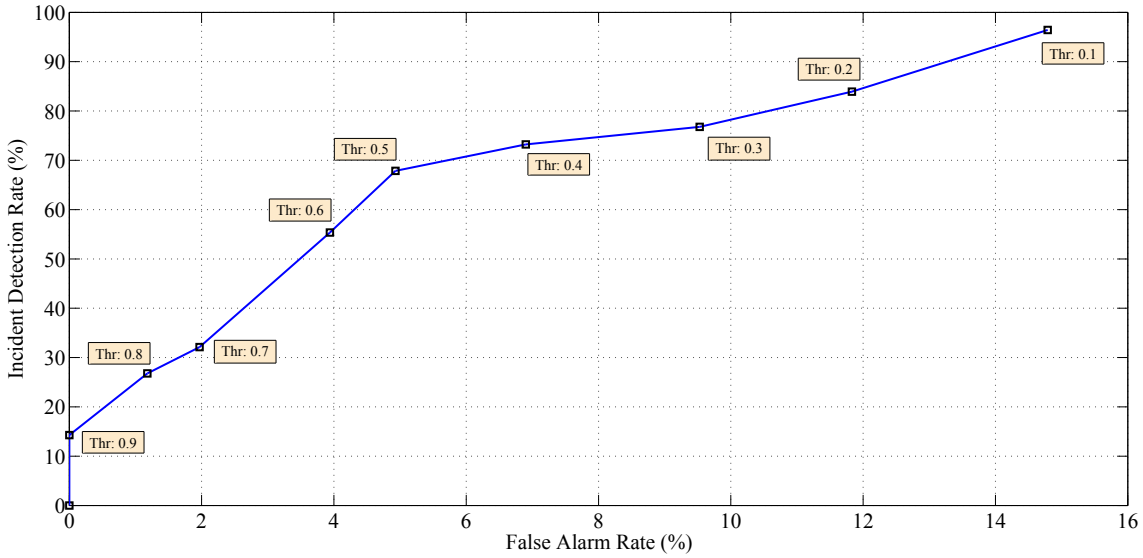

Figure 6.3: Incident Detection vs False Alarm Rate

6.6 Incident Detection Using DWT and Logistic Regression

6.6.1 Data Filtering

Data obtained through loop detectors has sometimes white noise added to it. For traffic analysis, prediction and control purposes, denoised and smoothened data is necessary. Wavelet transform techniques provide an effective way to denoise and have been applied successfully in various areas especially in image processing. In this research, we use wavelet transform to denoise and smoothen traffic data, obtained through loop detectors. This denoising technique is used as a preprocessing to the incident detection algorithm. Data obtained through this preprocessing is referred to as filtered data hereafter.

Noise is a random error that gets added to the observed data. In can be due to instrumental errors, technology limitations, human factor, or due to natural phenomenon such as atmospheric disturbances. Denoising algorithms try to separate the original signal and the additive white noise. Wavelet denoising, transforms the data into wavelet basis by decomposing it into the wavelet and scaling coefficients. Scaling coefficients, also called approximation coefficients, are large in values, whereas wavelet coefficients are very small in magnitude and represent variations in the signal. By appropriately choosing the threshold and applying them on the wavelet coefficients, we get rid of the noise. Denoised signal can be obtained back by inverse wavelet transforms using approximation and thresholded wavelet coefficients. Main steps of wavelet based denoising are as follows:

- Apply wavelet transform to the data and decompose it into approximation and

detail coefficients.

- Select an appropriate threshold and apply the thresholllding (either soft or hard depending on the data and objective)

- Inverse wavelet transform using approximation and thresholded wavelet coefficients.

Threshold Selection

There are two types of thresholding:

- Hard thresholding (keep or kill)

$$Thr = \begin{cases} median(abs(detail\ at\ level\ 1)); & if\ nonzero \\ 0.05\ max(abs(detail\ at\ level\ 1)); & otherwise \end{cases} \tag{6.9}$$

  In hard thresholding, the coefficients below a certain threshold are set to zero and the magnitudes of the wavelet coefficients above the threshold are left unchanged.

$$T_d^{hard} = \begin{cases} d\ ; & |d| > Thr \\ 0\ ; & |d| \leq Thr \end{cases}$$

- Soft thresholding (shrink or kill)

$$Thr = \sqrt{2.log(n)} \tag{6.10}$$

$$\text{and}\ \ n = prod(size(x))$$

In soft thresholding, the coefficients below a certain threshold are set to zero whereas the remaining coefficients are reduced by an amount equal to the value of the threshold.

$$T_d^{soft} = \begin{cases} sgn(d)(|d| - Thr) \ ; & |d| > Thr \\ 0 \ ; & |d| \leq Thr \end{cases}$$

Hard thresholding is default for compression whereas soft thresholding is recommended for denoising of a given signal.

6.6.2 Framework and Methodology

Traffic data is first filtered using DWT and high resolution components are discarded. Low resolution components are sufficient to represent the traffic flow data as explained in chapter 5.

For DWT of traffic signal we need to find out the best wavelet and optimum decomposition level for incident detection. Table 6.8 shows the incident detection rate with various wavelets decomposed upto level 1 and using all three traffic parameters (Avg Speed + Volume + Occupancy).

Our analysis show that decomposition upto level one optimizes the detection, decomposition beyond level degrades the incident detection results substantially. Among various wavelets we observe that db-1(Haar) and sym-1 perform almost similarly. We choose db-1(Haar) for our analysis.

Wavelet Type	Incident Detection Rate (%)	False Alarm Rate (%)
Db1	66.67	5.12
Db2	55.34	8.34
Db3	33.33	7.14
Sym1	66.67	5.12
Sym2	55.34	7.14

Table 6.8: Incident Detection Results using various Wavelets

6.6.3 Effect of Data Filtering Using DWT

In order to visualize effects of data filtering using DWT on improving the incident detection, we plot raw as well as filtered data obtained from RTC-FAST. Figure 6.4 and 6.5 show the plots of occupancy vs traffic volume for both incident and incident free cases. It can be observed that after filtering using DWT, regions of incidents and non incidents are more separable. This helps the logistic regression models in classification of incidents and non incidents, hence improving the incident detection rate.

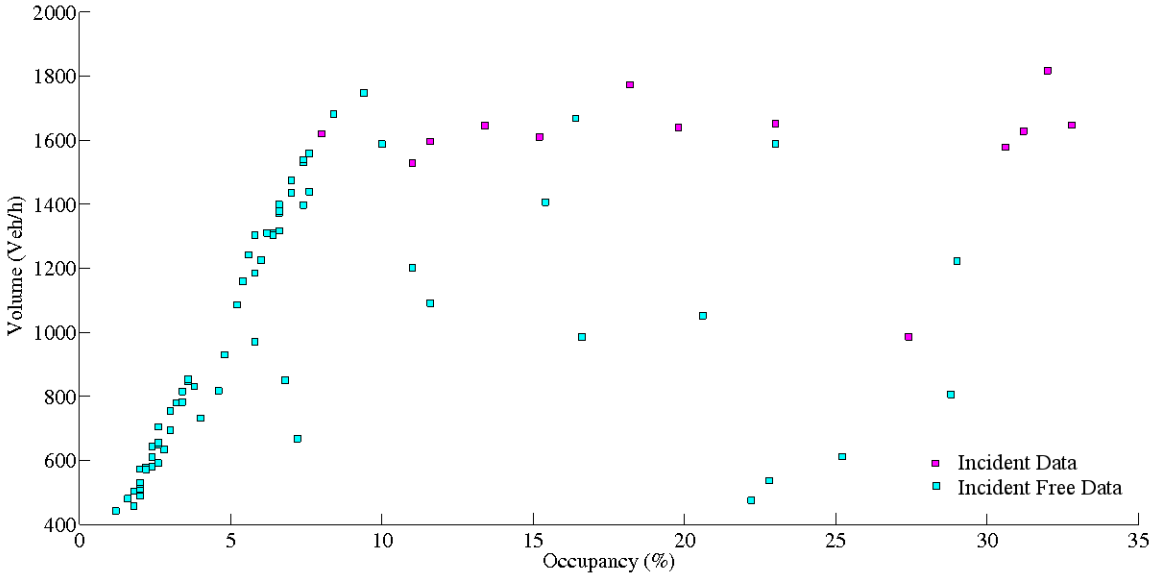

Figure 6.4: Raw Traffic Data

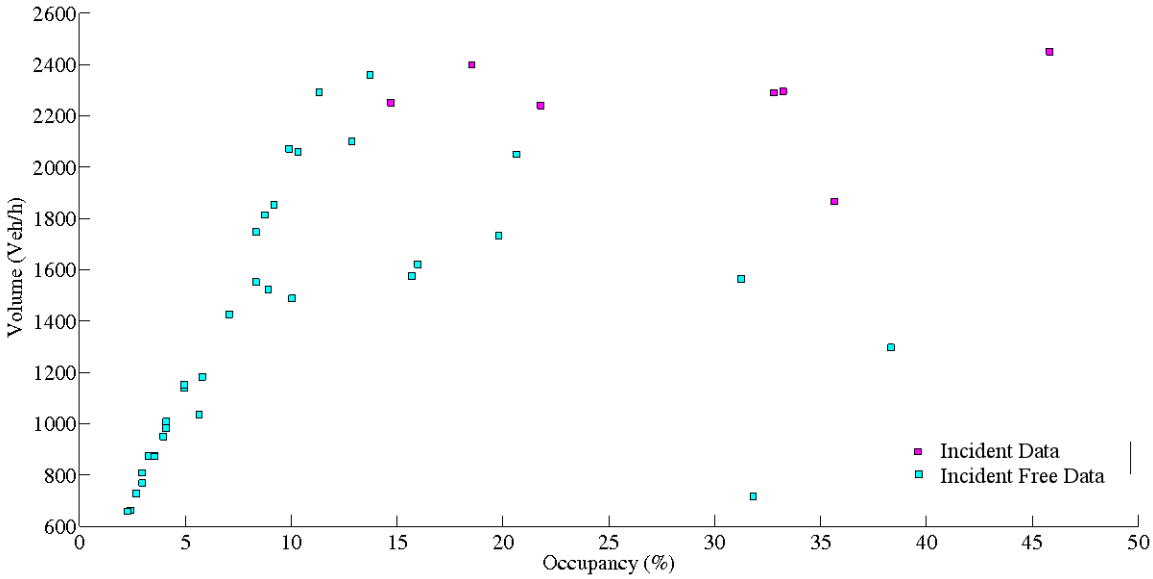

Figure 6.5: Filtered Traffic Data using DWT

6.6.4 Improvement in Incident detection using DWT

Table 6.9 shows incident detection results using logistic regression after preprocessing of the data by DWT.

Traffic Data	Incident Detection Rate (%)	False Alarm Rate (%)
Volume(Veh/h)	32.14	3.45
Occupancy (%)	60.71	6.90
Avg Speed (miles/h)	53.57	6.21
Avg Speed + Occupancy	71.43	4.14
Avg Speed + Volume	71.43	4.83
Occupancy + Volume	67.86	4.83
Avg Speed + Occupancy + Volume	75.00	4.14

Table 6.9: Incident Detection Results using DWT and Logit Models

It can be observed that the new hybrid model combining DWT and Logistic Regression, yields a better incident detection rate of 75% as compared to 64.15% using only logistic regression model. False alarm rate in this hybrid model is also on the lower side 4.14%, as compared to 5.26% in previous case. These results are for the combination all three traffic parameters i.e. Traffic Volume + Avg Speed + Occupancy. However as clear from Tables 6.7 and 6.9, for each of the parameters and their combinations, preprocessing the data using DWT yields in better detection rate and lesser false alarm.

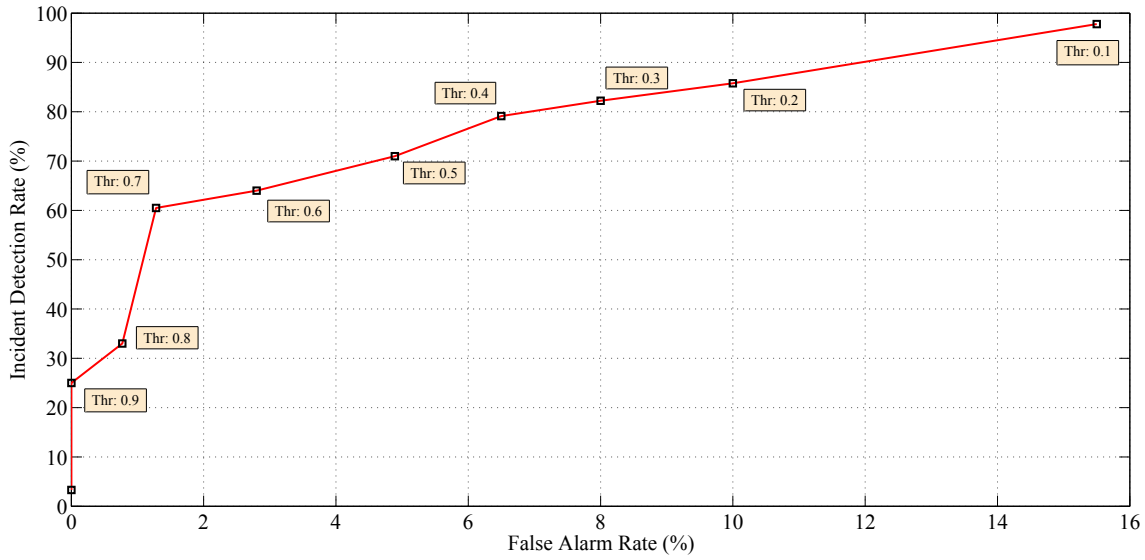

Figure 6.6: Incident Detection vs False Alarm Rate for filtered data

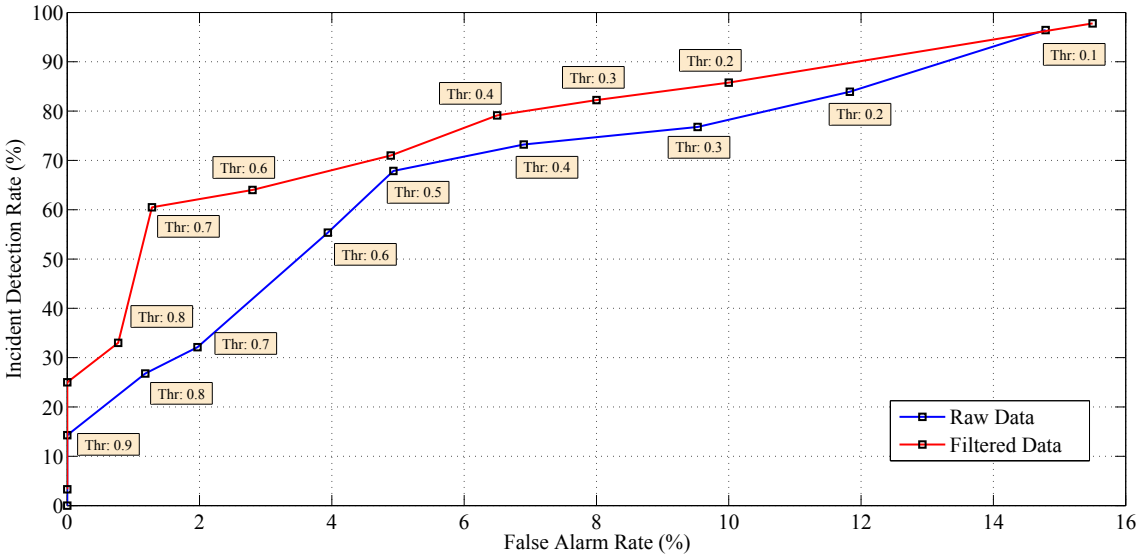

Figure 6.7: Comparison of Raw and Filtered Data

6.7 Conclusion

Logistic Regression technique was discussed and used for traffic incident detection. Various combinations of traffic parameters were tested and best detection rate of 64.15% (using 0.5 as default threshold) was observed for the combination of Traffic Volume + Occupancy + Avg Speed. Receiver Operating Characteristic (ROC) curves were plotted by varying the threshold, which showed a maximum of 95% detection rate for a 14% false alarm rate.

A new hybrid model was proposed combining two different computational approaches: wavelet transform and logistic regression. It was observed that detection rate improved with the new model (75%) and false alarm rate was also reduced to 4.14%. Receiver Operating Characteristic (ROC) curves were plotted by varying the threshold, and the curve was compared with the ROC curve of unfiltered data. It was observed that at each fixed false alarm rate, hybrid model gave a better incident detection rate.

CHAPTER 7

CONCLUSION AND FUTURE WORK

7.1 Conclusion

This thesis mainly focused on use of wavelets in Intelligent transportation systems. An introduction to Intelligent Transportation systems and Wavelet Theory was presented. Thesis was divided into two sections: Data compression and Incident Detection using wavelets.

First section focused on use of wavelet transform in transportation data compression. We investigated compression of loop detector data. We divided our analysis into three sections based on the data arrangement, 1D, 2D and 3D data arrangement. We explored EZW and SPIHT encoding schemes for compression of traffic data.

In the second section, Logistic Regression technique was discussed and used for traffic incident detection. Various combinations of traffic parameters were tested and best detection rate of 64.15% (using 0.5 as default threshold) was observed for the combination of Traffic Volume + Occupancy + Avg Speed. A new hybrid model was proposed combining two different computational approaches: wavelet transform and logistic regression. It was observed that detection rate improved with the new model (75%) and false alarm rate was also reduced to 4.14%. Receiver Operating Characteristic (ROC) curves were plotted by varying the threshold, and the curve

was compared with the ROC curve of unfiltered data. It was observed that at each fixed false alarm rate, hybrid model gave a better incident detection rate.

7.2 Future work

In the future, further investigation can be done on data compression techniques using wavelets. An algorithm can be developed which exploits features of both wavelet transform and traffic data. On the similar patters as image compression techniques (EZW and SPIHT), compression algorithm exploiting traffic data features, is expected to show better results.

Hybrid incident detection algorithm proposed can be further scaled for real time detection purposes. A cost analysis can be done of the false alarms, and balance can be struck between false alarm rate and incident detection rates which is economically and socially acceptable.

BIBLIOGRAPHY

[1] WISQARS, *Leading Causes of Death Reports*, (2007), <http://webappa.cdc.gov/sasweb/ncipc/leadcaus10.html>, (2011).

[2] *Injury Prevention & Control: Motor Vehicle Safety*, <http://www.cdc.gov/motorvehiclesafety/>, (07/10/2011).

[3] Arellano et al, Traffic Safety Analysis: A Spatially Enabled Technology, *ESRI* (2002).

[4] Souleyrette et al, Traffic Safety Analysis Software State of the Art, *Minnesota Department of Transportation* (February 2011).

[5] A Key Strategy to Optimize Surface Transportation System Performance, *4th Integrated Transportation Management Systems (ITMS) Conference*.

[6] Roger H.L. Chiang et al, Reverse engineering of relational databases: Extraction of an EER model from the relational database, *Data and Knowledge Engineering* ,12 (1994), 107-142.

[7] Roger H.L. Chiang et al, Knowledge-Based System for Performing Reverse Engineering of Relational Databases, *Decision Support Systems*, 13 (1995) 295-312 North-Holland.

[8] Roger H.L. Chiang et al, A Framework for The Design and Evaluation of Reverse Engineering Methods for Relational Databases, *Data & Knowledge Engineering*, 21 (1997) 57-77.

[9] Dowming Yeh et al, A Process for Extracting ER Diagram From a Table Based Legacy Database, *The Journal of Systems and Software*, 81 (2008) 764771.

[10] Ullman and Widom, *A First Course in Database Systems*, Prentice Hall, Upper Saddle River, New Jersey (2002).

[11] Hong, Wang and Gardner, *Real Analysis with an Introduction to Wavelets and Applications*, Elsevier Academic Press, Burlington, MA (2005).

[12] Adeli and Karim, *Wavelets in Intelligent Transportation Systems*, John Wiley & Sons Ltd., West Sussex, England (2005).

[13] Chang and Wang, Improved Freeway Incident Detection using Fuzzy Set Theory, *Transportation Research Part C: Emerging Technologies*,(Vol 3, Issue 6):371-388,(1995).

[14] Cheu and Ritchie, Automated detection of lane-blocking freeway incidents using artificial neural networks, *Transportation Research Board*,(1453):17-82,(1994).

[15] Adeli and Samant, An Adaptive Conjugate Gradient Neural NetworkWavelet Model for Traffic Incident Detection, *Computer-Aided Civil and Infrastructure Engineering*,(Vol 15, Issue 4):251-260,(2000).

[16] Samant and Adeli, Enhancing Neural Network Traffic Incident-Detection Algorithms Using Wavelets, *Computer-Aided Civil and Infrastructure Engineering*,(Vol 16, Issue 4):239-245,(2001).

[17] Rodrguez, G., *Lecture Notes on Generalized Linear Models*, (2007), <http://data.princeton.edu/wws509/notes/>,

[18] Ghosh, Basu et al, Analysis of trend in Vehicle Traffic data flow by Wavelets, *ISSC*,(2006).

[19] Daubechies, Ten Lectures on Wavelets, *CBMS-NSF Regional Conference Series in Applied Mathematics*

[20] Ding et al, A Method for Urban Traffic Data Compression Based on Wavelet-PCA, *Fourth International Joint Conference on Computational Sciences and Optimization*.

[21] Li et al, A Flow Volumes Data Compression Approach for Traffic Network Based on Principal Component Analysis, *Intelligent Transportation Systems Conference Seattle*,WA, USA, Sept. 30 - Oct. 3, 2007

[22] Torrence et al, A Practical Guide to Wavelet Analysis, *Bulletin of American Meteorological Society*,(Vol 79, Issue 1),(1998).

[23] Qiao, Liu and Yu, Incorporating Wavelet Decomposition Technique to Compress TransGuide Intelligent Transportation System Data, *Transportation Research Record*,1968,(2006).

[24] Qiao, Liu and Yu, Intelligent Transportation Systems Data Compression Using Wavelet Decomposition Technique, *Transportation Research Information Database*,(2009).

[25] Xiao-fa Shi, A Data Compression Method for Traffic Loop Detectors' Signals Based on Lifting Wavelet Transformation and Entropy Coding, *Information Science and Engineering (ISISE)International Symposium*,(2010).

[26] Jin and Ran, Automatic Freeway Incident Detection Based on Fundamental Diagrams of Traffic Flow, *Transportation Research Record: Journal of the Transportation Research Board*,(No. 2099):65-75, Washington DC, DOI: 10.3141/2099-08

[27] Yi, Sheng and Yu, Wavelet Transform for Feature Extraction to Improve Volume Adjustment Factors for Rural Roads, *Transportation Research Record: Journal of the Transportation Research Board*,(No. 1879):24-29, Washington DC, 2004

[28] Abdel-Aty, Uddin, Pande et al, Predicting Freeway Crashes from Loop Detector Data by Matched Case-Control Logistic Regression, *Transportation Research Record: Journal of the Transportation Research Board*,(No. 1897):88-95, Washington DC, 2004

[29] Brown, Foo, et al, Comparison and Analysis Tool for Automatic Incident Detection, *Transportation Research Record: Journal of the Transportation Research Board*,(No. 1925):58-65, Washington DC, 2005

VITA

Graduate College
University of Nevada, Las Vegas

Shaurya Agarwal

Home Address:
1861, Arbol Verde Way
Las Vegas, Nevada 89119

Degrees:
Bachelor of Technology, Electronics & Communications Engineering, 2009
Indian Institute of Technology, Guwahati, India

Master of Science, Electrical Engineering, 2012
University of Nevada, Las Vegas, NV

Thesis Title: Wavelets in Intelligent Transportation Systems

Thesis Examination Committee:
Co-Chair, Dr. Pushkin Kachroo
Co-Chair, Dr. Emma Regentova
Committee Member, Dr. Ke-Xun Sun
Graduate College Representative, Dr. Haroon Stephen