STATE OF CALIFORNIA • DEPARTMENT OF TRANSPORTATION

TECHNICAL REPORT DOCUMENTATION PAGE

TR0003 (REV 10/98)

ADA Notice

For individuals with sensory disabilities, this document is available in alternate formats. For information call (916) 654-6410 or TDD (916) 654-3880 or write Records and Forms Management, 1120 N Street, MS-89, Sacramento, CA 95814.

1. REPORT NUMBER CA16-2257	2. GOVERNMENT ASSOCIATION NUMBER	3. RECIPIENT'S CATALOG NUMBER
4. TITLE AND SUBTITLE Work Zone Injury Data Collection and Analysis		5. REPORT DATE 08-28-2015
		6. PERFORMING ORGANIZATION CODE AHMCT Research Center, UC Davis
7. AUTHOR Bahram Ravani, Patricia Fyhrie, Kristopher Wehage, Arash Gobal, and Hiu Y. Hong		8. PERFORMING ORGANIZATION REPORT NO. UCD-ARR-15-06-30-01
9. PERFORMING ORGANIZATION NAME AND ADDRESS AHMCT Research Center UCD Dept. of Mechanical & Aerospace Engineering Davis, California 95616-5294		10. WORK UNIT NUMBER
		11. CONTRACT OR GRANT NUMBER 65A0395, Task ID 2257
12. SPONSORING AGENCY AND ADDRESS California Department of Transportation Division of Research, Innovation, and System Information P.O. Box 942873 Sacramento, CA 94273-0001		13. TYPE OF REPORT AND PERIOD COVERED Final Report June 2011 – June 2015
		14. SPONSORING AGENCY CODE Caltrans

15. SUPPLEMENTARY NOTES

16. ABSTRACT

Work-zone related injuries and fatalities are a major safety concern in California and nationwide. Developing mitigation measures is vital in improving work zone safety both for roadway workers as well as the traveling public. Developing such measures, however, require detailed data on the characteristics of these accidents and injuries produced in them as well as injury costs models that can be used for cost benefit assessments. Although there exist databases and data sources such as the Statewide Integrated Traffic Records Systems (SWITRS), NHTSA's (National Highway Traffic Safety Administration's) FARS (Fatality Analysis Repotting System) database, and Caltrans TASAS (Traffic Accident Surveillance and Analysis System), none can provide the information that would justify particular mitigation measures, or allow cost benefit analysis. This research was conducted to develop such data, codify and classify it in terms of factors and outcomes and provide analysis tools in terms of injury costs. It involved collecting, codifying and classifying all Traffic Collison Reports for accidents occurring near a work-zone from 12 Caltrans districts for a period of five years (2006-2010). Extracted data from these reports were codified in terms of factors and outcomes and made part of a decision support system designed to allow analysis of the data that can be used for planning and management of work-zone operations to improve worker and motorist safety.

17. KEY WORDS Work Zone Accidents, Fatalities, Injuries, Cost benefit Analysis	18. DISTRIBUTION STATEMENT No restrictions. This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161.	
19. SECURITY CLASSIFICATION *(of this report)* Unclassified	20. NUMBER OF PAGES 62	21. COST OF REPORT CHARGED

Reproduction of completed page authorized

DISCLAIMER/DISCLOSURE STATEMENT

The research reported herein was performed as part of the Advanced Highway Maintenance and Construction Technology (AHMCT) Research Center, within the Department of Mechanical and Aerospace Engineering at the University of California – Davis, and the Division of Research, Innovation and System Information at the California Department of Transportation. It is evolutionary and voluntary. It is a cooperative venture of local, State and Federal governments and universities.

This document is disseminated in the interest of information exchange. The contents of this report reflect the views of the authors who are responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the State of California, the Federal Highway Administration, or the University of California. This publication does not constitute a standard, specification or regulation. This report does not constitute an endorsement of any product described herein.

For individuals with sensory disabilities, this document is available in Braille, large print, audiocassette, or compact disk. To obtain a copy of this document in one of these alternate formats, please contact: the Division of Research, Innovation and System Information, MS-83, California Department of Transportation, P.O. Box 942873, Sacramento, CA 94273-0001.

Advanced Highway Maintenance and Construction Technology Research Center

Department of Mechanical and Aerospace Engineering
University of California at Davis

Work Zone Injury Data Collection and Analysis

Bahram Ravani: Principal Investigator
&
Patricia Fyhrie, Kristopher Wehage, Arash Gobal, and Hiu Y. Hong

Report Number: CA 15-2257
AHMCT Research Report: UCD-ARR-15-06-30-01
Final Report of Contract: 65A0395, Task ID 2257

August 18, 2015

California Department of Transportation

Division of Research, Innovation and System Information

ABSTRACT

Work-zone related injuries and fatalities are a major safety concern in California and nationwide. Developing mitigation measures is vital in improving work zone safety both for roadway workers as well as the traveling public. Developing such measures, however, require detailed data on the characteristics of these accidents and injuries produced in them as well as injury costs models that can be used for cost benefit assessments. Although there exist databases and data sources such as the Statewide Integrated Traffic Records Systems (SWITRS), NHTSA's (National Highway Traffic Safety Administration's) FARS (Fatality Analysis Reporting System) database, and Caltrans TASAS (Traffic Accident Surveillance and Analysis System), none can provide the information that would justify particular mitigation measures, or allow cost benefit analysis. This research was conducted to develop such data, codify and classify it in terms of factors and outcomes and provide analysis tools in terms of injury costs. It involved collecting, codifying and classifying all Traffic Collison Reports for accidents occurring near a work-zone from 12 Caltrans districts for a period of five years (2006-2010). Extracted data from these reports were codified in terms of factors and outcomes and made part of a decision support system designed to allow analysis of the data that can be used for planning and management of work-zone operations to improve worker and motorist safety.

TABLE OF CONTENTS

Contents

Abstract ..... *ii*
Table of Contents ..... *iii*
List of Figures ..... *v*
List of Tables ..... *vi*
List of Acronyms and Abbreviations ..... *vii*
Acknowledgments ..... *viii*
Executive Summary ..... *ix*
Background ..... **ix**
Approach ..... **ix**
Results, Recommendations, and Limitations ..... **x**
Results x
Recommendations xiii
Limitations xiv
Chapter 1 ..... *1*
Introduction *1*
Problem **1**
Background **2**
Research Approach **2**
Overview of Research Results and Benefits **4**
Chapter 2 ..... *5*
data collection *5*
Identification of Traffic Collision Reports **5**
Collecting the Collision Report Files **9**
Semi-Automated Redaction Process **10**
Chapter 3 ..... *14*
Data Extraction and processing *14*
Chapter 4 ..... *16*
Work Zone Injury Decision Support Tool *16*
Decision Support Tool **16**
Using the search functionality 18
Results Page 21
Displaying Report Details 23

DATA ANALYSIS ***26***

Bibliography ***31***

Appendix A ***32***

Automatic Traffic Collision Report Optical Text Recognition Analyzer **32**

Open-Source Dependency **32**

Tesseract **32**

PyEnchant **33**

Spelling Corrector **33**

PyPDFOCR **34**

Google Services **34**

Direction API **34**

Tesseract Training Tools **35**

Training Procedure 35

Editing Box Files 36

Traffic Collision Report Processing Program **37**

Text correction 37

Collision Information Extraction 38

Installation/Configuration **44**

System Requirement 44

Installation 45

Execution 45

Known Bugs 45

LIST OF FIGURES

Figure 1. AHMCT Work zone Accident Decision Support Tool. x
Figure 2. Average Yearly Distribution of Accidents by Injuries in Numbers. xi
Figure 3. Average Yearly Distribution of Accident by Cost. xii
Figure 4. Average Yearly Distribution of Accidents by Cause. xiii
Figure 5 Typical first three pages of a TCR which is also referred to as "Form 555". 6
Figure 6 Screenshots of portions of the text file that were provided by the TASAS data query. 9
Figure 7. Example of results from redaction of personal information from the first three pages of a TCR. 11
Figure 8. Block Diagram of the Redaction Process. 12
Figure 9. Pull Down Menu for Redaction. 12
Figure 10. Pull-down Menu for Saving a Redacted TCR. 13
Figure 11. The Process of Data Extraction and Populating the Data Base. 16
Figure 12. The URL for the Decision Support Tool. 16
Figure 13. Architectural Diagram of the Decision Support Tool. 17
Figure 14. Welcome Page of the Decision Support Tool Web Site. 18
Figure 15. Search Page of the Decision Support Tool. 19
Figure 16. Searching By Categories. 21
Figure 17. Reports Matching Search Criteria. 22
Figure 18. Mapping of All Searched Reports to the Accident Locations. 23
Figure 19. Accident Report Page. 24
Figure 20. Output of a PC-CRASH Simulation. 25
Figure 21- Distribution of work zone collision costs for the time period from 2006 through 2010. 27
Figure 22. Total work zone collision cost for Daylight only collisions. 28
Figure 23. Total work zone collision cost for Nighttime only collisions. 28
Figure 24- Number of collisions averaged per year shown as a percentage of totals. 29
Figure 25- Pie chart of collision frequency per primary cause. 30
Figure 26. Pie chart of collision frequency per primary cause. 30
Figure 27. Extracting the collision data from the first page of TCR. 39
Figure 28- Portions highlighted from Page 1 of a TCR where basic collision information is located. 40
Figure 29. Highlighted region in a TCR indicates where data and time are located 40
Figure 30- Highlighted region in a TCR indicates where date and time are located on "all the other pages". 41
Figure 31- First page of a TCR with party information highlighted. 41
Figure 32- Vehicle information is highlighted on the first page of the Traffic Collision Report. Note the vehicle information is always associated with a party number. 42
Figure 33- The region where the driver information is highlighted is illustrated here. 42
Figure 34- Region on the collision report where vehicle damage is designated. 43
Figure 35- Towed or driven away for a particular vehicle is highlighted here. 43
Figure 36- Highlighted area shows where victim injury is described in detail. 44

LIST OF TABLES

Table 1. Distribution of work-zone accidents by year for Caltrans districts.....ix
Table 2 List of field/variable names in the TASAS database.....7
Table 3 Number of TASAS reports for years 2006-2010 distributed amongst Caltrans Districts for "Road Condition=D".....8
Table 4. Number of TCRs successfully tracked down and scanned and available through the AHMCT Injury Database.....10
Table 5- Cost Model based on crash severity.....26
Table 6- Total Work Zone Collision Costs broken up by year and Caltrans District.....27

LIST OF ACRONYMS AND ABBREVIATIONS

AHMCT	Advanced Highway Maintenance and Construction Technology Research Center
API	Application Programming Interface
Caltrans	California Department of Transportation
CHP	California Highway Patrol
DOT	Department of Transportation
DRISI	Caltrans Division of Research, Innovation and System Information
IT	Information Technology
MAIS	Maximum Abbreviated Injury Scale
MAIT	CHP Multidisciplinary Accident Investigation Teams
OSHA	Occupational Safety and Health Administration
PDF	Publishable Document Format
PDO	Property Damage Only
QA/QC	Quality Assurance / Quality Control
SHSP	Strategic Highway Safety Program
SR	State Route
SWITRS	Statewide Integrated Traffic Records System
TASAS	Traffic Accident Surveillance and Analysis System
TCRPP	Traffic Collision Report Processing Program
TCR	Traffic Collision Report (CHP Form 555)
TIF	Tagged Image Format
TIFF	Tagged Image Format File
WZSAFETY	Work Zone Safety web site

ACKNOWLEDGMENTS

The authors would like to thank the California Department of Transportation (Caltrans) for the support of this research. Special thanks go to Hamid Ikram and Juan Araya of Caltrans without their contributions this project could not have succeeded. The authors also acknowledge the dedicated efforts many undergraduate students who participated in data collection and coding. In addition the help from many individuals at various Caltrans offices are greatly appreciated. These include:

- Headquarters: Debbie Silva (now retired) and Hamid Ikram
- District 1: Marsha Davenport**,** Ralph Marinelli
  (Their office was in an upheaval due to an office renovation. Years2006 and2007 were boxed up and unable to be pulled out of storage at the time of data collection)
- District 2: Gerry Reyes
- District 3: Truc Nguyen, Darryl Chamber, Lori Marino
- District 4: Kapsoon Capulong
- District 5: Steve Cadenasso
- District 6: Dan Singh
- District 7: Seyed Torabzadeh
- District 8" Andrew Machen, Julie Grifen
- District 9: Greg Weirick and Johnny Bhullar who had visited the site on other Caltrans business and did us the favor of getting copies of the reports for us
- District 10: Duane Hawkes, Duper Tong
- District 11: Charles Gray
- District 12: Bryan Sorensen

EXECUTIVE SUMMARY

Background

Work-zone related injuries and fatalities are a major safety concern in California and nationwide. According to the FHWA (Federal Highway Administration) one work zone-related injury occurs every 14 minutes and one work-zone related fatality occurs every 15 hours resulting in 96 injuries and 1.6 fatalities a day. Existing estimates suggest that work zone accidents and injuries have direct medical costs of over $800 million per year.

Although there exist databases and data sources such as the Statewide Integrated Traffic Records Systems (SWITRS) based upon California Highway Patrol crash reports, NHTSA's FARS database or OSHA databases, and Caltrans TASAS (Traffic Accident Surveillance and Analysis System), none can provide the information that would justify particular mitigation measures, because they report only outcomes and locations. For mitigation purposes, however, much more information is needed on the resulting property damage, severity of injuries, and methods to estimate associated costs. In addition, information about the collision in terms of "what hit what", and contributing factors related to the causation of the accidents can be very useful. This research was conducted to develop such data, codify and classify it and provide a decision support type analysis tool.

Approach

This research collected, codified, and classified all Police Traffic Collison Reports for accidents occurring near a work-zone from 12 Caltrans districts for a period of five years (2006-2010). The distribution of such accidents by year and Caltrans districts is shown in Table below.

Table 1. Distribution of work-zone accidents by year for Caltrans districts.

Year/ District	2006	2007	2008	2009	2010	total
1	31	76	56	44	41	248
2	80	56	68	70	74	348
3	104	234	314	488	395	1535
4	607	710	713	821	607	3458
5	220	259	297	194	146	1116
6	182	293	171	173	119	938
7	940	885	1112	712	729	4378
8	1587	1268	675	510	622	4662
9	9	2	8	16	7	42
10	276	348	323	211	104	1262
11	489	406	320	327	364	1906
12	1077	617	303	255	210	2462
Total	5602	5154	4360	3821	3418	22355

Extracted data from these reports were codified in terms of factors and outcomes and made part of a decision support tool designed to allow analysis of the data that can be used for planning and management of work-zone operations to improve worker and motorist safety.

Results, Recommendations, and Limitations

Results

This research has resulted in a comprehensive data set on work-zone related accidents in California for a five year period. The data set is configured into a decision support tool (Figure 1) that can be used for planning and work zone management and other safety evaluation purposes. This tool allows cost benefit analysis of different work zone configurations based on a cost model developed that can be indexed with the consumer price index.

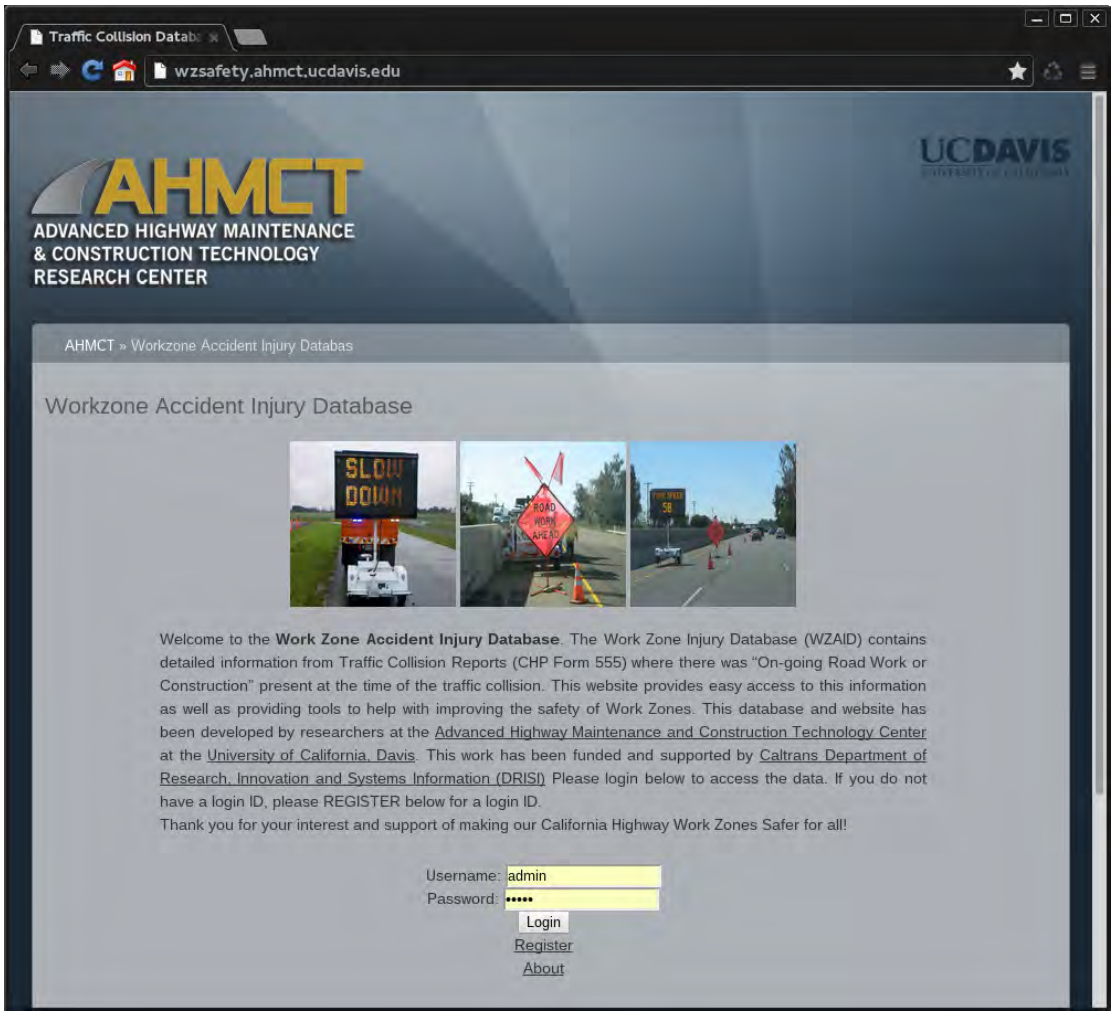


Figure 1. AHMCT Work zone Accident Decision Support Tool.

The decision support tool can be used to query data and analysis related to work zone accidents. For example, if one is interested to understand the average percentage of accidents resulting in

Property Damage Only (PDO) versus those with injuries and fatalities the data can be retried and plotted in a pie chart as depicted in Figure 2. Data in this Figure indicates that majority of work zone accidents (in numbers) involve property damage only with accidents involving fatalities or severe injuries constituting less than approximately 2% of all work zone accidents.

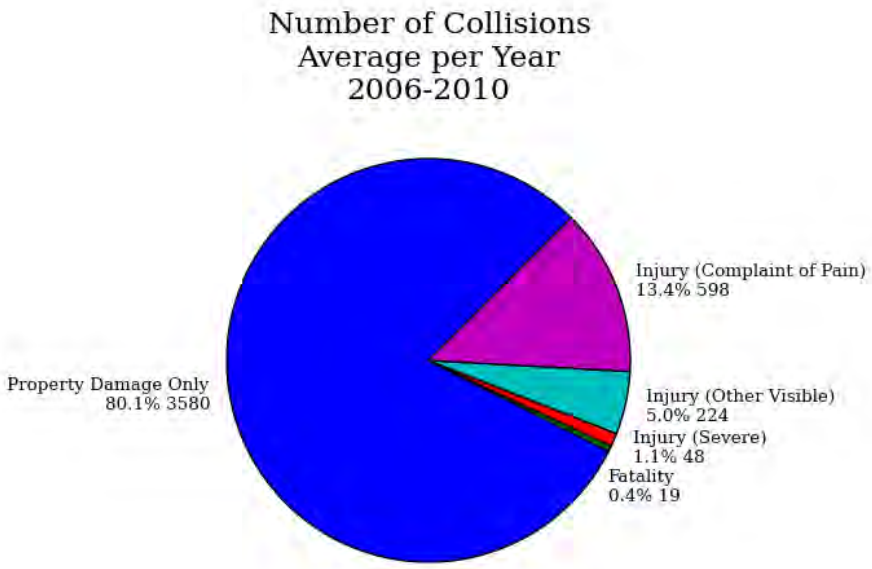


Figure 2. Average Yearly Distribution of Accidents by Injuries in Numbers.

The decision support tool developed as part of this research also has injury cost models that can be used to estimate the cost of the outcomes of accidents in terms of the level of injuries in them. For example, the yearly average cost of different types of work zone accidents are calculated by the decision support tool and plotted in a pie chart as depicted in Figure 3.

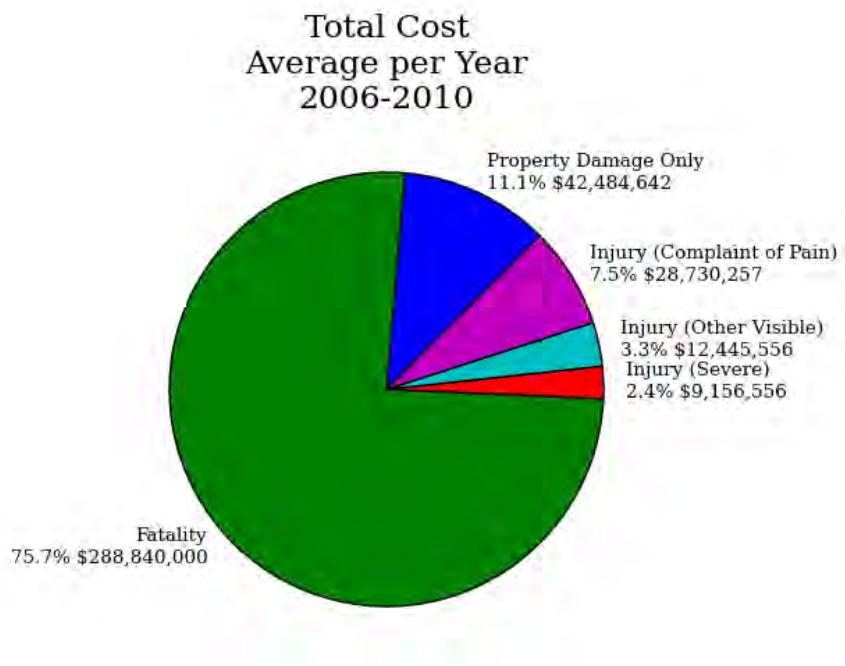


Figure 3. Average Yearly Distribution of Accident by Cost.

It is clear from data in Figure 3, however, that the major costs of work zone accidents are related to accidents involving fatalities and sever injuries. The yearly average cost of all work zone collisions is approximately of $382 Million. This is much lower than estimates that existed before this research study indicating these costs being in excess of $800 Million per year. The cost estimates provided here are based on actual data for California over a five year period rather than being based on speculations.

The decision support tool allows evaluation and assessment of causes of work zone accidents. The data on average yearly distribution of accidents by their main causes is depicted in the pie chart of Figure 4.

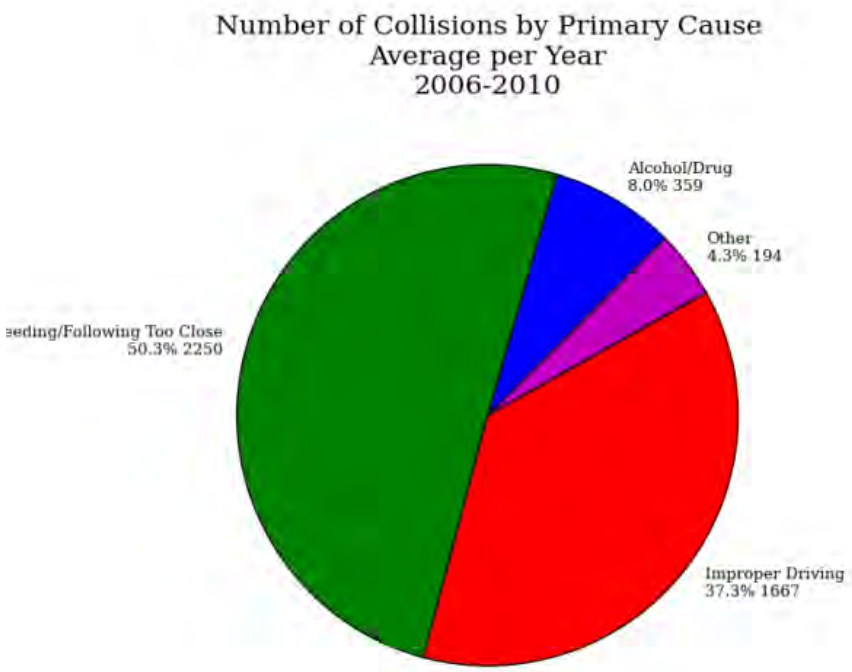


Figure 4. Average Yearly Distribution of Accidents by Cause.

It is clear from Figure 4, that speeding, following too closely, and improper driving constitutes the major portion of work zone accidents. This means that mitigation measure addressing driver behavior may have the best outcome in improving work zone safety.

Recommendations

The recommendations for the use of the decision support tool or the data set includes:

- Using injury and fatality cost models for budgeting of COZEEP/MAZEEP operations.
- Evaluation of hot spots in terms of accidents in planning and management of highway operations.
- Cost-benefit analysis of using different barrier systems for different work zone configurations and locations.
- Evaluating accident severity to determine the right level of attenuators for Truck Mounted and other Attenuators.
- Using the automated data extraction method and scanning to make a business change in Caltrans district operations for collecting accident data such that it would feed the decision support tool.

- Develop further research to perform computer simulation of most of the work zone accidents to identify all causes and better understand accident avoidance and reduction scenarios.
- Develop research to integrate the decision support tool with Caltrans Earth.
- Develop a Feasibility Study Report (FSR) will allow full deployment of the decision support tool in Caltrans.

The decision support tool and the data set developed in this research has already been used to address several safety and planning questions that has come up by different Caltrans units. It has also provided data for cost benefit analysis and budgeting for COZEEP/MAZEEP operations as well as supporting Challenge Areas 14 efforts. It can be used remotely by Caltrans users through the AHMCT web site.

Limitations

In using the data and analyses tools in the decision support tool, one has to keep in mind the several limitations including:

- The data is only for years 2006-2010.

- Cost models are based on cost of injuries and value of life established by Caltrans at the time of this research but can be updated using updates in the consumer price index.

- Injury data is based on limited information in the Police Collision Reports and may not be consistent with what can be found in actual medical records of the injured parties.

- The property damage data is also based on limited data in Police Traffic Collision Report and may not be accurate.

CHAPTER 1
INTRODUCTION

Work zone accidents and injuries are a major safety concern and data is needed to understand the nature and causes of these and their economic and social impacts so that mitigation measures can be considered and developed. Estimates suggest that work zone accidents and injuries cost over $800 million per year but there is no real data to back this up scientifically. In addition, there are costs associated with property damage, lost earnings, lost household production, travel delay, vocational rehabilitation, workplace costs, administrative costs, legal costs, pain, and lost quality of life. According to the FHWA (Federal Highway Administration) one work zone-related injury occurs every 14 minutes and one work-zone related fatality occurs every 15 hours resulting in 96 injuries and 1.6 fatalities a day. It is therefore important to evaluate actual accident data to understand the frequency and the nature of work zone accidents and develop cost models that would provide a basis for consideration and justification of different mitigation methods. This research study was conducted to collect detailed data on work zone accidents that occurred on California Highways for a period of five years. The objective was to have a complete data set that would allow planners, decision makers to consider in addressing different issues and to make their decisions and assessments data driven rather than based on partial and incomplete information.

Problem

Improving safety along California highway's work zone sites is a notable component of California's Strategic Highway Safety Program (SHSP) [1]. Since each work zone site is relatively unique with respect to configuration, number of open/closed lanes, presence of cones or barriers, etc. In order to improve safety and reduce the risk/severity of collisions, a thorough description of the cause(s) and location of the collision relative to the work zone are critical pieces of information.

Although there exist databases and data sources such as the Statewide Integrated Traffic Records Systems (SWITRS) based upon California Highway Patrol crash reports, NHTSA's FARS database or OSHA (Occupational Safety and Health Administration) databases, and Caltrans TASAS (Traffic Accident Surveillance and Analysis System), none can provide the information that would justify particular mitigation measures, because they report outcomes and locations, but not information such as driver behavior, work zone intrusion, work zone location, number of lanes, and comments by drivers, witnesses, and officers. TASAS provides basic outcome information such as how many people were hurt or killed what was the basic event that took place (e.g. auto accident, car hitting the barrier, etc.)? For mitigation purposes, however, much more information is needed. These include data on the nature and severity of injuries, methods to estimate medical costs associated with the injuries, more information about the collision in terms of "what hit what", localized information about the actual location in the work zone where the accident occurred (taper, activity zone or transition area), and finally more information about contributing factors related to the causation of the accidents. All such information is not

included in TASAS and can play crucial role in developing and planning for mitigation measures.

This research study was aimed at collecting detailed data that can be used for evaluation of causes of accidents and injuries so that it can be analyzed to evaluate mitigation strategies. The research was also intended to develop injury cost models so that some of the economic impact of work zone accidents can be quantified. The research involved collecting data for a five year period for all accidents that occurred near a work zone from its police Traffic Collision Report (TCR) for all Caltrans districts. This data was codified and was combined with injury cost models in a data base that can be used for analysis and other evaluation purposes. The results obtained are a step forward in satisfying the needs of California's SHSP.

Background

As part of a project related to performing cost benefit analysis of a mobile highway barrier system, AHMCT researchers collected partial data on work zone accidents from three Caltrans districts. This involved evaluating the full text of California Highway Patrol (CHP) 555 Traffic Collision Reports for a two year period – a total of 2,370 Traffic Collision reports. Later on AHMCT researchers studied the details of these collected reports and synchronized the data with 18,100 Caltrans TASAS records. This provided data and a methodology for a powerful tool for safety planning and risk reduction for work zone and highway safety. For example, in our detailed analysis, we first identified 40 distinct causal factors and 19 distinct outcomes for work zone accidents. We then summarized these factors into 7 basic causes and outcomes. Next we looked closely at injury descriptions and categorized each incident in terms of the Abbreviated Injury Scale (AIS) developed by the Association for Advancement of Automotive Medicine.

An analysis of CHP Traffic Collision Reports (TCR) is labor intensive and complicated. Lack of resources is most likely the reason it has not been done in the past. Also until very recently, only paper copies of collision reports were available along with the fact these reports are distributed throughout all of California. What was performed in this research study was to expand and complete the work begun under another study by including data from all 12 Caltrans districts and for a span of five years. This allowed a better understanding of nature, cause, and cost of injuries in work zone accidents as well as an understanding of the effect of different highway corridors on accident and injury frequencies. It also allowed adding other important information to the data set that can potentially facilitate work zone planning. The results also provide a statically valid set of data for analysis related to work zone injuries and accidents.

Research Approach

The AHMCT Research Center at the University of California-Davis proposed to gather data throughout all of California on work zone collisions and respective injuries. Following the data collection, the data was then organized in a form useful for planning and assessing work zone safety situations and developing mitigation measures. At present, existing data to justify any particular mitigation measure is based on conjecture. Using current traffic collision reporting databases such as TASAS and SWITRS do not provide any information on the work zone and its role or its' contribution to the vehicle collision. The one place where detailed work zone

collision data does exist however is in the traffic collision report itself written up by the attending CHP officer. The officer reports on the physical attributes of the crash site, where the collision occurred with respect to the work site, names and statements of those involved, determination of causes, along with other factors which influenced the collision.

The approach for this project involves obtaining and "processing" the original CHP TCR (also known as Form 555) document on record for all accidents involving work zones in all 12 Caltrans districts and over a period of five year. The overall work flow of the project followed these basic steps:

1. Form the Technical Advisory Group (TAG)
2. Develop Base Line Data Parameters and Needed forms for Data Collection
3. Identify Liaison persons at Caltrans Districts
4. Data Collection
5. Data Integrity Analysis
6. Data Coding and Analysis
7. Documentation and Presentation of Research Results to Caltrans

To support the "Data Collection" phase (Step 4 above), a "template" was developed to systematically redact all personal information such as names, addresses, and birthdates. Each TCR is then redacted and stored in a database. In addition, relevant detailed information from each TCR was extracted, codified and stored in the database where it can be analyzed.

The approach for data collection and coding consisted of the following steps:

1. Paper TCRs from all 12 Caltrans district offices which were identified as "work zone related", retrieved and scanned into a *.pdf format file.
2. All personal information was redacted from each *.pdf file with the help of the "template" developed as part of this research.
3. A work zone collision database was developed and was first populated with associated TASAS data and SWITRS data where appropriate.
4. The following information (when available) was then gathered from all the TCRs:
    a) Traveling speed of involved vehicles.
    b) The relative position of the collision site with respect to the work zone.
    c) Whether there was intrusion into the work zone itself. This also includes hitting any positive protection item or construction equipment.
    d) Detailed information of the injuries
    e) Age of the driver
5. Each collision diagram was then saved in its own separate *.jpg file.
6. All narrative information in a TCR was then codified from an image file to a searchable *.pdf file with searchable, recognized text.
7. The injury database was populated with all this information.

Overview of Research Results and Benefits

This research resulted in collection of data for work zone accidents from all 12 Caltrans districts for the years 2006-2010. Data from a total of over 20,000 accidents that occurred in California work zones during this period were collected and codified and stored in a database for future analysis.

The benefits of this research include the data and analysis results that would allow responses to at least the following questions:

a. What is the nature and severity of work zone accidents?
b. What factors, outcomes and attributes are important in terms of injuries and fatalities?
c. What are the factors that affect causation of work zone accidents?
d. What are estimates of injury costs and what factors influence injury severity?

Such data with proper analysis and simulation can provide the basis for evaluating different mitigation strategies and will result in improvement of highway safety for both highway workers and the traveling public.

CHAPTER 2

DATA COLLECTION

This section discusses the data collection methodology used in this research. The first step, in this process, involved identification of specific TCRs which were related to a collision at or near a work zone in California. Then the TCR had to be collected and all the personal information on it had to be redacted. The data and pertinent information in the TCR was then codified and made into an electronic format and put into a database. These steps are described in the remainder of this section.

Identification of Traffic Collision Reports

When a collision occurs on a Caltrans maintained highway, it is tracked by means of the Traffic Accident Surveillance and Analysis System" (TASAS). This is a database maintained by Caltrans where key information is stored regarding the location, date, time, along with external conditions. For a collision to be entered into this database, the collision must occur on a state highway (maintained by Caltrans), involve at least 1 motorized vehicle, can involve any injury level (no injury to any injury to fatality), and documented by a CHP officer. The CHP officer enters the collision information into a Form 555 otherwise known as a TCR.

A copy of the TCR can be seen in Figure 5. The first page typically gives the basic information related to the collision; the second page provides special "coding" variables that both TASAS and the SWITRS databases rely upon. If there are any injuries or fatalities, then the third page is included. The third page can also be used to capture witness information even if there are no injuries. After these three pages are typed, then the reporting officer typically includes a schematic diagram of the collision's physical description. The narrative portion of the report is then attached. All report pages will contain the same header information.

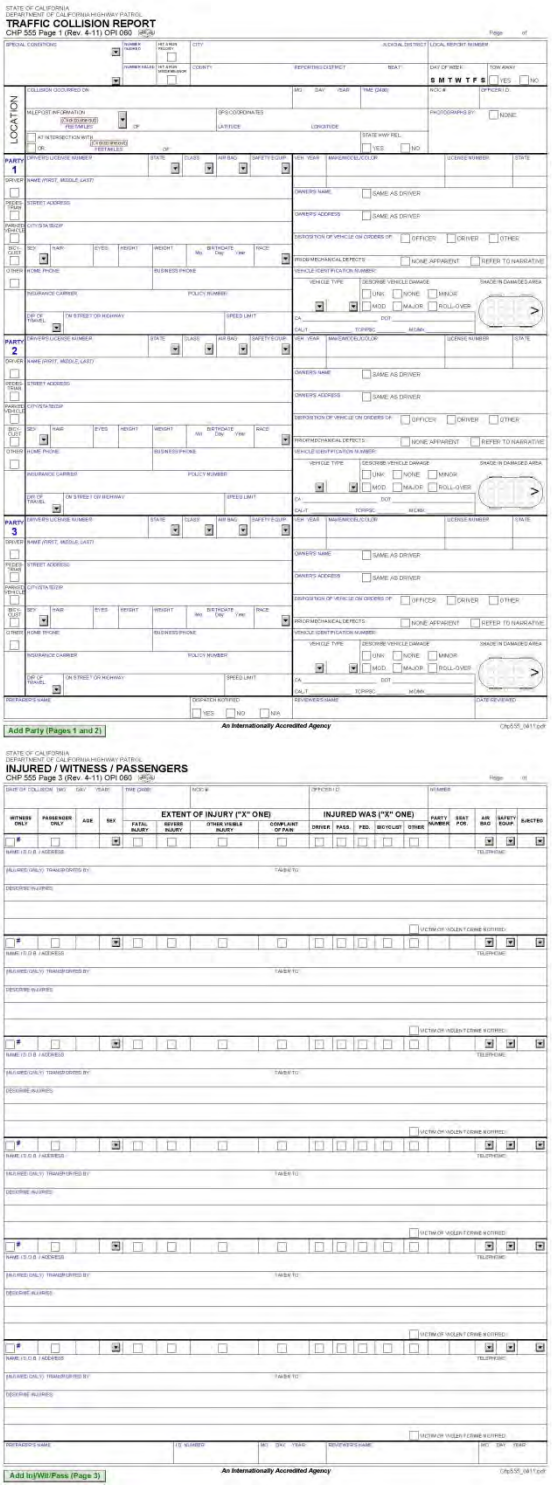


Figure 5 Typical first three pages of a TCR which is also referred to as "Form 555".

At the time of this data collection, there was a well established protocol (developed earlier by AHMCT) where targetted fields in a TCR were extracted and inserted into the TASAS database. This procedure is thoroughly described in Chapter 3 of the Caltrans Traffic Manual (Chapter 3: Accident and Roadway Records). A list of all variables available in the TASAS database can be seen in Table 2. In TASAS all reports can be uniquely identified with: Date, Time, County, Highway, and Post Mile Information. With respect to the overall size of his database, there are typeically over 150,000 collisions a year on state highways so this particular database is quite large.

Table 2 List of field/variable names in the TASAS database.

Year District Route County Post Mile Highway Group Access Control Median Type Barrier Type Number Of Lanes Left Number Of Lanes Right Population Code File Type Intersection / Ramp Accident Location	Side Of Highway Day Of Week Accident Date Accident Time Accident Number Primary Collision Factor Weather Lighting Roadway Surface Roadway Condition Right Of Way Control Type Of Collision Number Of Motor Vehicles Involved Party Type	Direction Of Travel Vehicle Highway Indicator Special Information Persons Killed Persons Injured Primary Object Struck Location A Other Object Struck Location B Other Object Struck Location C Other Object Struck Location Other Associated Factor Movement Preceding Collision Location Sobriety Drug Physical

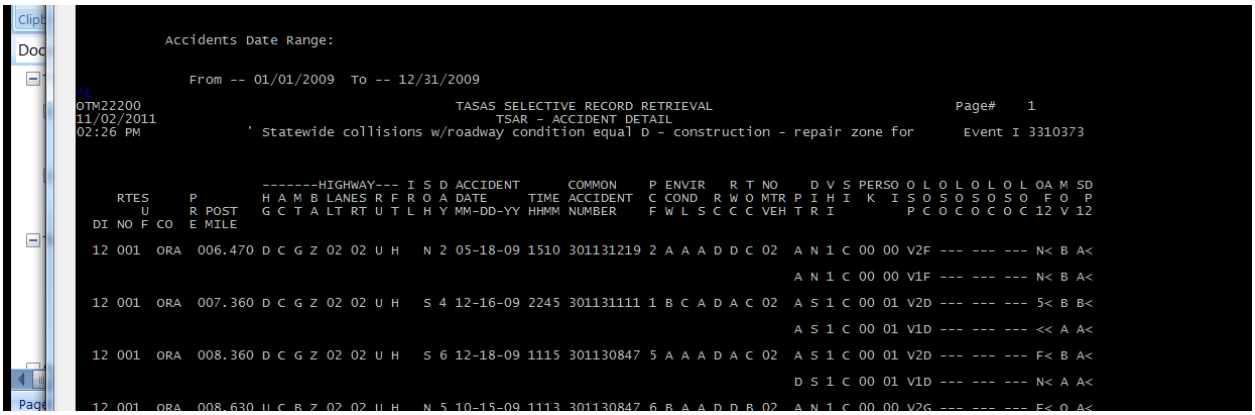
In data collection part of this research in order to discern amongst all of these collision reports to find which ones took place in a construction or maintenance work zone, the TASAS variable name “Road Condition” was used. When the variable “Road Condition = D” (Ongoing Road Work), it was assumed the collision occurred adjacent to the ongoing road work site or preceding it if lane closures were present at the time of the collision.

To extract the list of work zone collisions from the state’s TASAS database, a “Query” was requested to the headquarters TASAS personnel to gather the records where “Road Condition = D” and Year=2006, 2007, 2008, 2009, and 2010. The number of reports for each District and each year that resulted from this query can be seen in Table 3.

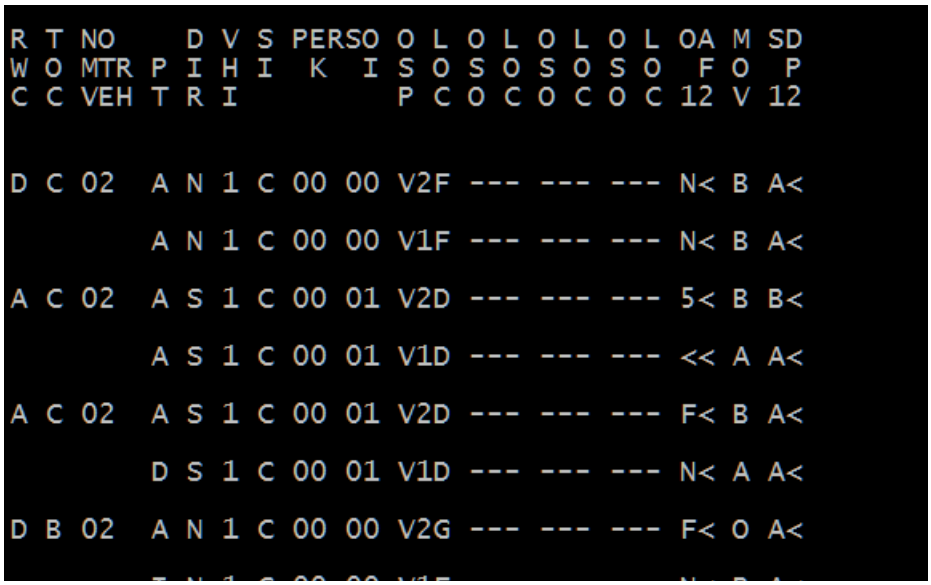
Table 3 Number of TASAS reports for years 2006-2010 distributed amongst Caltrans Districts for "Road Condition=D".

District	2006	2007	2008	2009	2010
1	31	76	56	44	41
2	80	56	68	70	74
3	104	234	314	488	395
4	607	710	713	821	607
5	220	259	297	194	146
6	182	293	171	173	119
7	940	885	1112	712	729
8	1587	1268	675	510	622
9	9	2	8	16	7
10	276	348	323	211	104
11	489	406	320	327	364
12	1077	617	303	255	210
Total	5602	5154	4360	3821	3418
	Total for all 5 years = 22,355				

The text files that were provided were printer output from a mainframe-style database. An example of the data can be seen in Figure 6. In this research a “Python” script had to be written to remove the header information and titles, page numbers, etc. from these files. The result was a list of work zone collision record information with which the needed information was available to go to Caltrans District offices and perform the next task. As a side note, the TCR data was sorted by Caltrans district number with the quantities reflected in Table 3.



```
          Accidents Date Range:

             From -- 01/01/2009  To -- 12/31/2009

OTM22200                                   TASAS SELECTIVE RECORD RETRIEVAL                          Page#    1
11/02/2011                                      TSAR - ACCIDENT DETAIL
02:26 PM            ' Statewide collisions w/roadway condition equal D - construction - repair zone for      Event I 3310373


                    -------HIGHWAY--- I S D ACCIDENT      COMMON    P ENVIR   R T NO    D V S PERSO O L O L O L O L OA M SD
    RTES     P      H A M B LANES R F R O A DATE     TIME ACCIDENT  C COND  R W O MTR P I H I  K  I S O S O S O S O  F O  P
       U     R POST G C T A LT RT U T L H Y MM-DD-YY HHMM NUMBER    F W L S C C C VEH T R I        P C O C O C O C 12 V 12
 DI NO F CO  E MILE

 12 001  ORA  006.470 D C G Z 02 02 U H   N 2 05-18-09 1510 301131219 2 A A A D D C 02  A N 1 C 00 00 V2F --- --- --- N< B A<

                                                                                        A N 1 C 00 00 V1F --- --- --- N< B A<

 12 001  ORA  007.360 D C G Z 02 02 U H   S 4 12-16-09 2245 301131111 1 B C A D A C 02  A S 1 C 00 01 V2D --- --- --- 5< B B<

                                                                                        A S 1 C 00 01 V1D --- --- --- << A A<

 12 001  ORA  008.360 D C G Z 02 02 U H   S 6 12-18-09 1115 301130847 5 A A A D A C 02  A S 1 C 00 01 V2D --- --- --- F< B A<

                                                                                        D S 1 C 00 01 V1D --- --- --- N< A A<

 12 001  ORA  008.630 U C B Z 02 02 U H   N 5 10-15-09 1113 301130847 6 B A A D D B 02  A N 1 C 00 00 V2G --- --- --- F< O A<
```



```
R T NO    D V S PERSO O L O L O L O L OA M SD
W O MTR P I H I  K  I S O S O S O S O  F O  P
C C VEH T R I        P C O C O C O C 12 V 12

D C 02  A N 1 C 00 00 V2F --- --- --- N< B A<

        A N 1 C 00 00 V1F --- --- --- N< B A<

A C 02  A S 1 C 00 01 V2D --- --- --- 5< B B<

        A S 1 C 00 01 V1D --- --- --- << A A<

A C 02  A S 1 C 00 01 V2D --- --- --- F< B A<

        D S 1 C 00 01 V1D --- --- --- N< A A<

D B 02  A N 1 C 00 00 V2G --- --- --- F< O A<
```

Figure 6 Screenshots of portions of the text file that were provided by the TASAS data query.

Collecting the Collision Report Files

Once the TASAS output from TCRs at a work zone was generated for years 2006-2010, twelve subsets of this information were generated. For each of the 12 Caltrans Districts, a "Key Info" list was generated for the data-collection researchers to "track down" the TCR hardcopy from the file room. Each TCR is identified as a unique document by the following "Key Info" data:

- County
- Highway number and direction (N, S, E or W)
- Post mile revision code and post mile marker of collision locations
- Data
- Time

Each Caltrans District was visited by 1-3 data-collection researchers between Jan 2012-November 2013. Depending on the number of TCRs to track down, the amount of time spent at a Caltrans district office was varied. Once at the Caltrans office and acquiring secure access to the TCR File Room, the data-collection researchers set up laptops and scanners to scan, redact personal information, save and encrypt the pdf files output from the scanning process. On the average, a researcher would process 250-300 files in an 8 hour day. This included set-up time, pulling the files, un-stapling the TCRs, scanning, redacting, storing the data, re-stapling and re-filing the TCR.

There were issues however that caused the final count of scanned reports to be less than that indicated by the TASAS query output (Table 3). Reasons why a TCR may be missing from the final collection included situations in which the original TCR was not filed correctly, not seen or "checked out". Other times, the image quality from the resultant TCR scan had inadequate quality for further processing. The final numbers of TCRs collected and codified in this research are shown in Table 3

Table 4. Number of TCRs successfully tracked down and scanned and available through the AHMCT Injury Database.

District	2006	2007	2008	2009	2010
1	0	0	56	43	31
2	76	51	58	63	72
3	31	73	100	483	312
4	597	682	703	808	602
5	104	196	295	188	140
6	162	254	147	160	107
7	845	858	1008	665	479
8	1037	773	630	477	558
9	9	2	8	16	5
10	3	204	312	200	61
11	110	44	314	306	145
12	1036	236	286	236	194
Total	4010	3373	3917	3645	2706
	acquired for all 5 years = 17,651				

Semi-Automated Redaction Process

In collecting data as part of this research study, it was important to redact all personal and identifying information from each TCR before it was scanned into the data base of the decision support tool developed in this research study. The first three pages of TCR had data that needed

to be redacted. Since each TCR was scanned into a pdf format, a program in JavaScript was written in Adobe Acrobat as an add-on tool that would redact all the fields with personal and identifying information in a pdf of a TCR. An example of what the redacted pages look like can be seen in Figure 7.

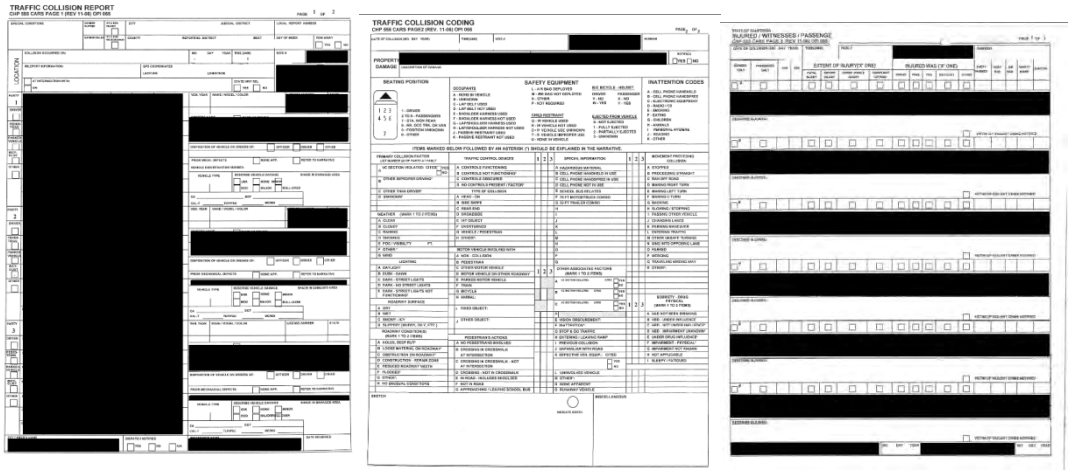


Figure 7. Example of results from redaction of personal information from the first three pages of a TCR.

Since the quality of pdf scans of TCRs were varied due to misalignment of pages on the scanner or existence of other unrelated marks on the page, the redaction process required human interaction. The process is shown in a block diagram in Figure 8. There were two points in the process that there was a need for human interaction. The process involved selecting a proper template for redaction of information and then checking the quality of the results until all the information was clearly redacted. A total of four templates were developed based on the variety in the quality of the scanned pdf files. In the future if the TCRs will be created electronically then the redaction process can be fully automatic.

The redaction process consisted of first using a pull down menu in Adobe Acrobat for redaction of a TCR as depicted in Figure 9. The user would then have to start with a template and check the results to see if the redaction was complete and accurate. The add-on-tool would then allow the user to again use the pull down menu to save the redacted file properly coded with the Caltrans district and the year of the accident. This is shown in Figure 10.

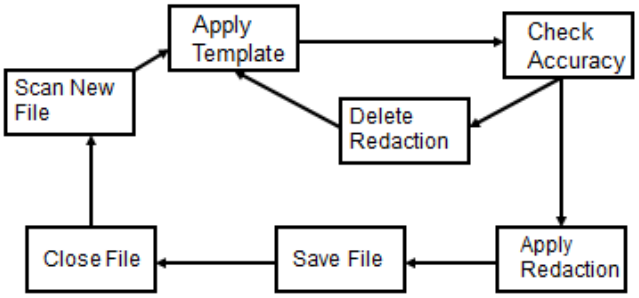


Figure 8. Block Diagram of the Redaction Process.

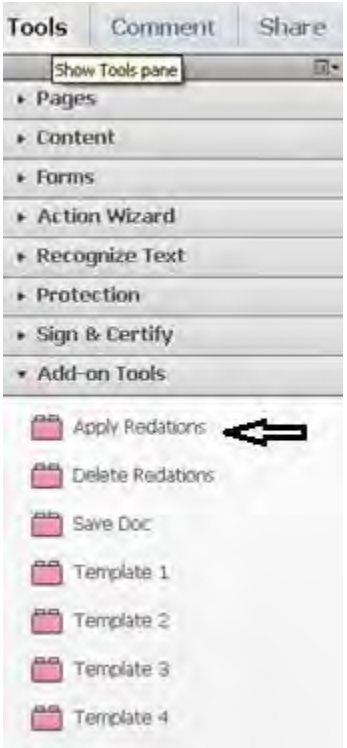


Figure 9. Pull Down Menu for Redaction.

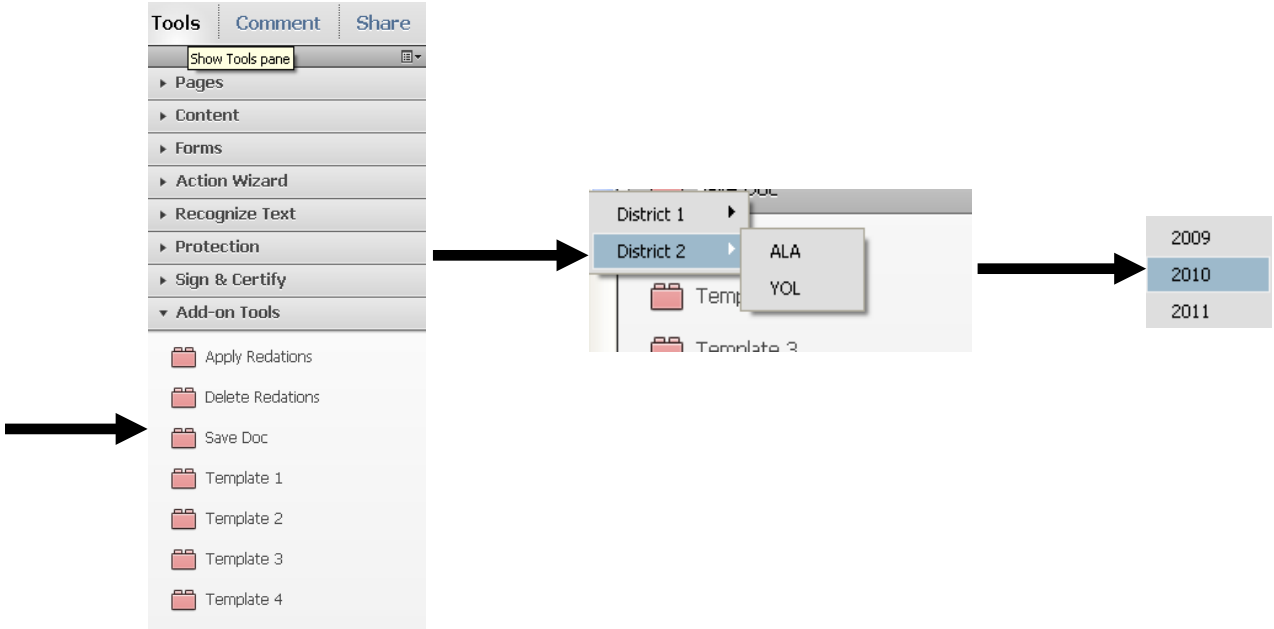


Figure 10. Pull-down Menu for Saving a Redacted TCR.

CHAPTER 3
DATA EXTRACTION AND PROCESSING

This research study had two major components: one was the collection and redaction of TCRs from all Caltrans districts for a period of five years and the second was extracting work zone safety related information from each redacted TCR. The following work zone safety related information was extracted from each TCR:

1. ***Injury Details***
   For each person injured or killed, there is typically detailed information on the injury such as what type of injury occurred and which body region was injured. In SWITRS a numbering system is used to rank the severity of the collision (1=Fatality, 2= Serious injury, 3= other visible injury, 4= Complaint of pain, 0=No injury). A great deal of insight can be obtained, however, if instead biomechanical principles are applied to determine impact speeds, collision angles, force levels and acceleration levels. This is what was done in this study.

2. ***Crash Severity Details***
   The crash severity information was based on the level of injury as well as whether or not the vehicle was towed away. If the damage to the vehicle was such that it was towed away that would indicate a higher severity as compared to the vehicle being driven away. In evaluating severity based on injury, the same coding system used in SWITRS was utilized. In SWITRS injury severity is coded as follows: 1=Fatality, 2= Serious injury, 3= other visible injury, 4= Complaint of pain, 0=No injury.

3. ***Driver's Age***

4. ***Type of Vehicle***

5. ***Traveling Speed***
   This is different than the posted speed limit. This value was taken from the narrative portion of the TCR from the statements of the involved parties. The data on traveling speed can therefore not be accurate since it is based on the statement of parties involved.

6. ***Whether or not there was Intrusion into the work zone***

In addition the collision schematics were taken and electronically added to the data set for each accident and all the narrative pages were coded electronically in searchable text format. All these were done to increase the future utility of the data set allowing users to search for different

key words or to be able to see collision diagrams to reconstruct the accidents and simulate the effects of different mitigation scenarios on the outcome.

Due to the very large number of TCRs that needed to be processed so that relevant data can be extracted and coded into a data base, Optical Character Recognition (OCR) technology was utilized to assist in extraction of information from some of the TCRs. The OCR technology developed and used in part of this research study is described in more detail in Appendix A. It should be pointed out that the use of OCR technology allowed finishing the coding of all the TCRs within the time period of this research study.

CHAPTER 4
WORK ZONE INJURY DECISION SUPPORT TOOL

In order to provide easy access to the data collected as part of this research study, a database was developed and was populated with the data extracted from the TCRs as well as the redacted forms of all the TCRs collected. The process of data extraction from a TCR to populating the data base is shown schematically in Figure 11. As is depicted in this figure, the scanned file of a redacted TCR consists of multiple pages that were processed either by a researcher or by the OCR software. The data in the form of optically recognizable characters were then processed and codified and dumped into a data base referred to here as Work Zone Injury data base.

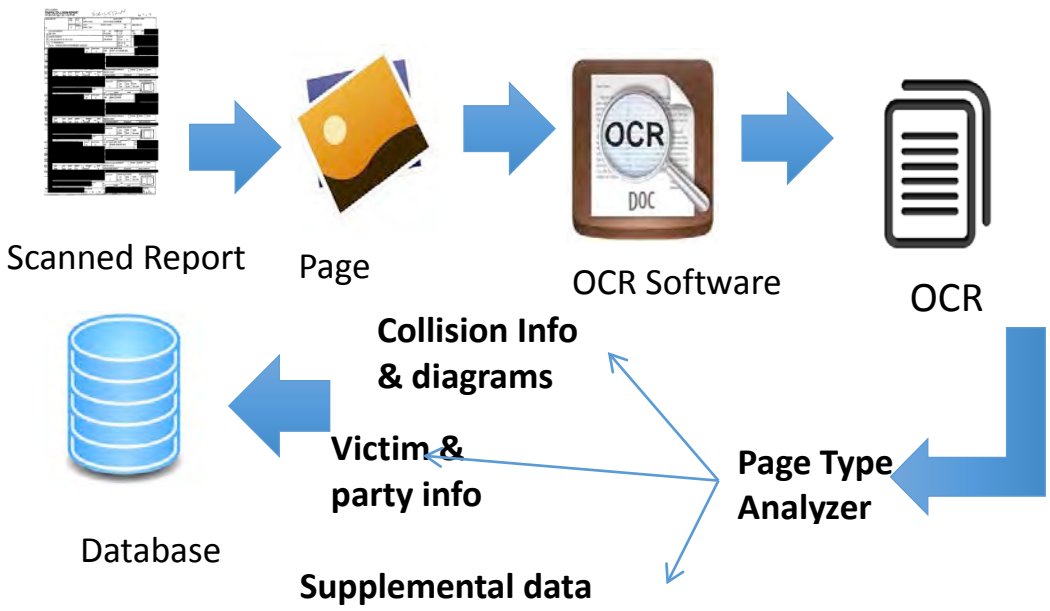


Figure 11. The Process of Data Extraction and Populating the Data Base.

Decision Support Tool

In order the improve the usability of the Work Zone Injury data base, a decision support tool with a web interface was designed and implemented around the Work Zone Injury data base. This decision support tool can be accessed through Work Zone Safety web site (WZSAFETY) with the URL address depicted in Figure 12.

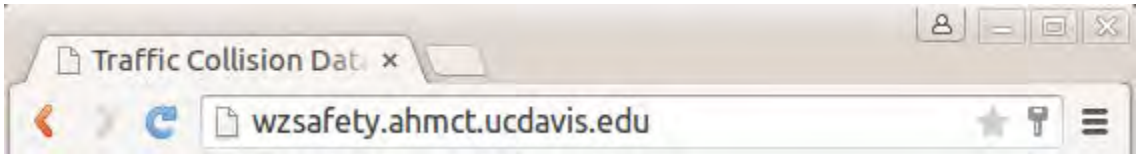


Figure 12. The URL for the Decision Support Tool.

Such a development was not part of the scope of this research study but it was undertaken to increase the usability of the work zone injury data base and facilitate its utilization in practice.

An architectural diagram of the decision support tool is depicted in Figure 13. As it can be seen in this figure, the decision support tool consists of the database that has a programming interface allowing its future expansions and modifications, a web framework and it is offered over a proxy server with a graphical user interface designed to facilitate its utility. The decision support tool provides easy access to the data stored in the work zone accident injury database. Registered users can search for collisions having special attributes and view all the details of selected collisions as well as download redacted TCRs.

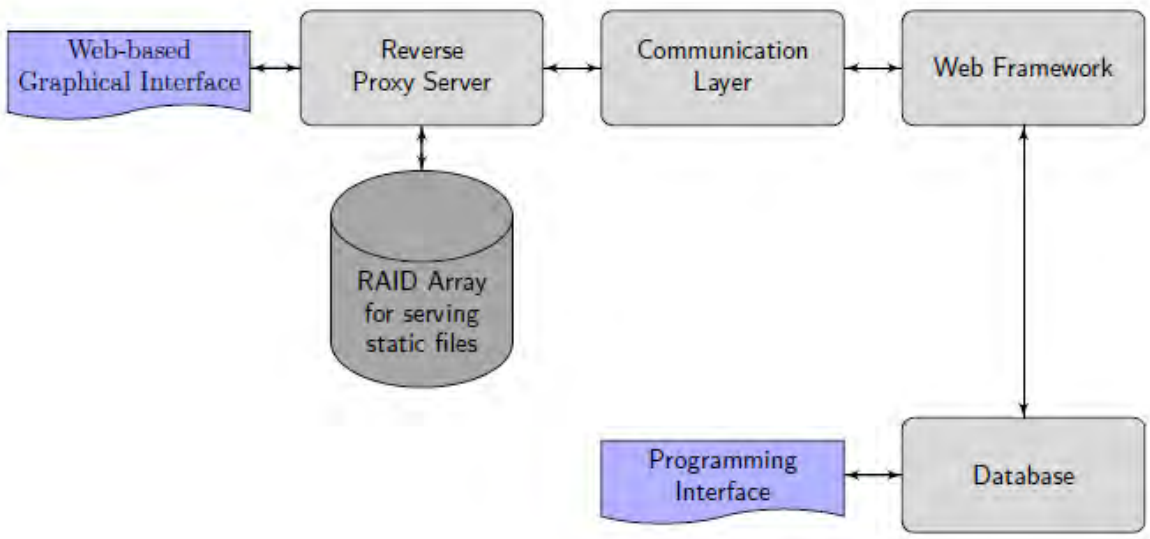


Figure 13. Architectural Diagram of the Decision Support Tool.

The WZSAFETY website is a secure web-based application based on a current web framework. The user must have an authenticated set of username and password to access the website. Accessing the web site requires use of Chrome 4 or higher version as a web browser. When the web site is accessed, a welcome page, as shown in Figure 14, appears. The user has to type in her/his username and password. After logging in, the website redirects the user to the search page that can be used to filter the reports or the data from the reports. This is described in more detail in the next section.

Figure 14. Welcome Page of the Decision Support Tool Web Site.

.

Using the search functionality

After logging in, the website redirects the user to the search page shown in Figure 15. Here, one can filter the reports by the following categories:

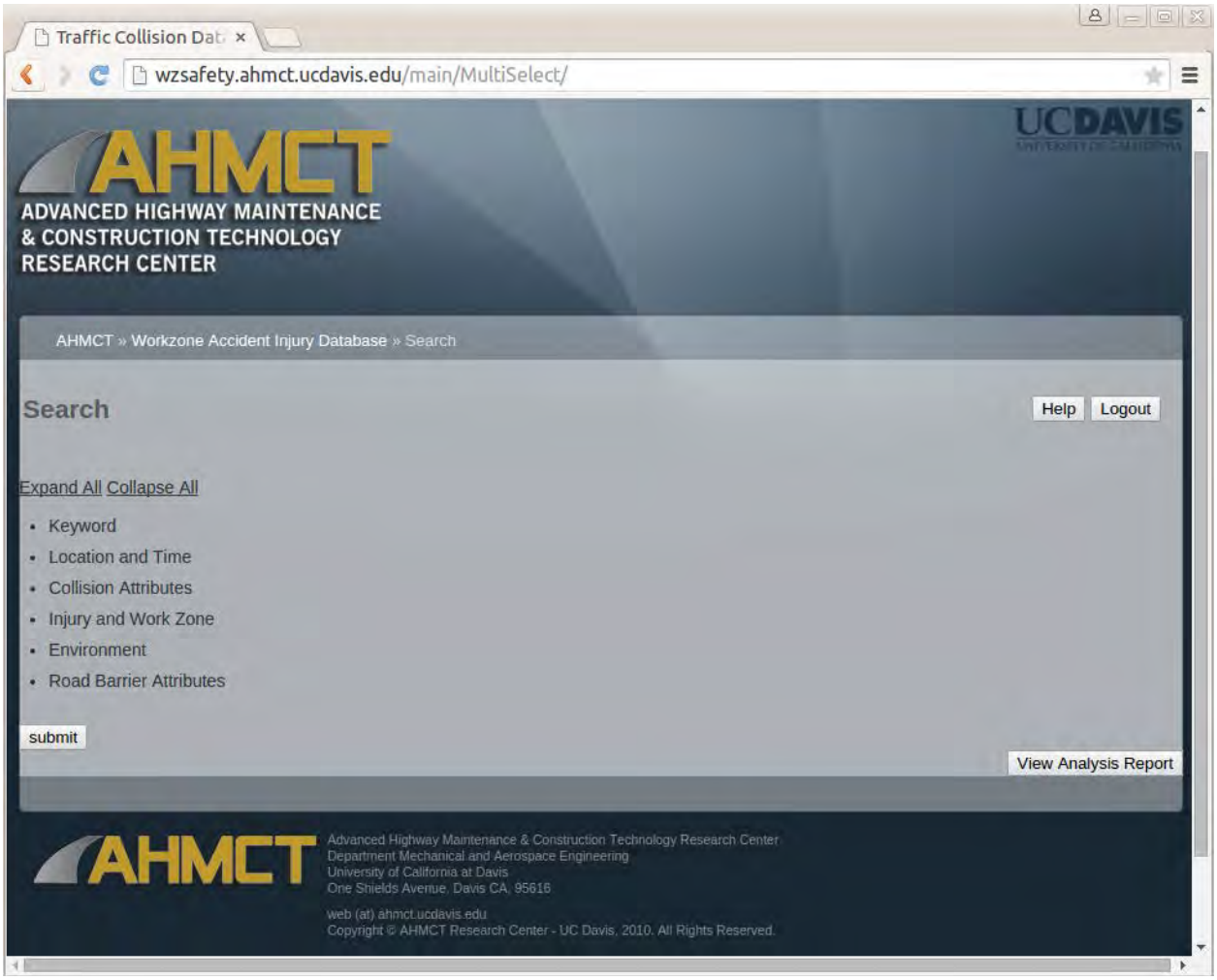


Figure 15. Search Page of the Decision Support Tool.

- Keyword
  - Shows only the reports having the keyword in their OCR text
- Location and time
  - Between two specific dates
  - Year
  - Day of week
  - Caltrans district number
  - County
  - Route number
- Collision Attributes
  - Type of collision (Head-on, Rear end etc.)
  - Primary cause of collision
  - Tow away

    - Number of involved parties
    - TASAS party type
    - SWITRS party type
    - Range of driver age
    - range of travelling speed
- Injury and Work Zone
    - Number of injuries
    - Number of fatalities
    - Intrusion into work zone
- Environment
    - Weather conditions
    - Lighting conditions
    - Population code (Urban, rural, etc.)
- Road attributes
    - Access type
    - Highway type
    - Caltrans road type
    - Highway side
    - Road surface conditions
    - Number of lanes
    - Barrier type
    - Median type

One can clip on each category name to display the sub-categories within it. One can also click on the expand-all button to show all subcategories. The category search function is illustrated in Figure 16.

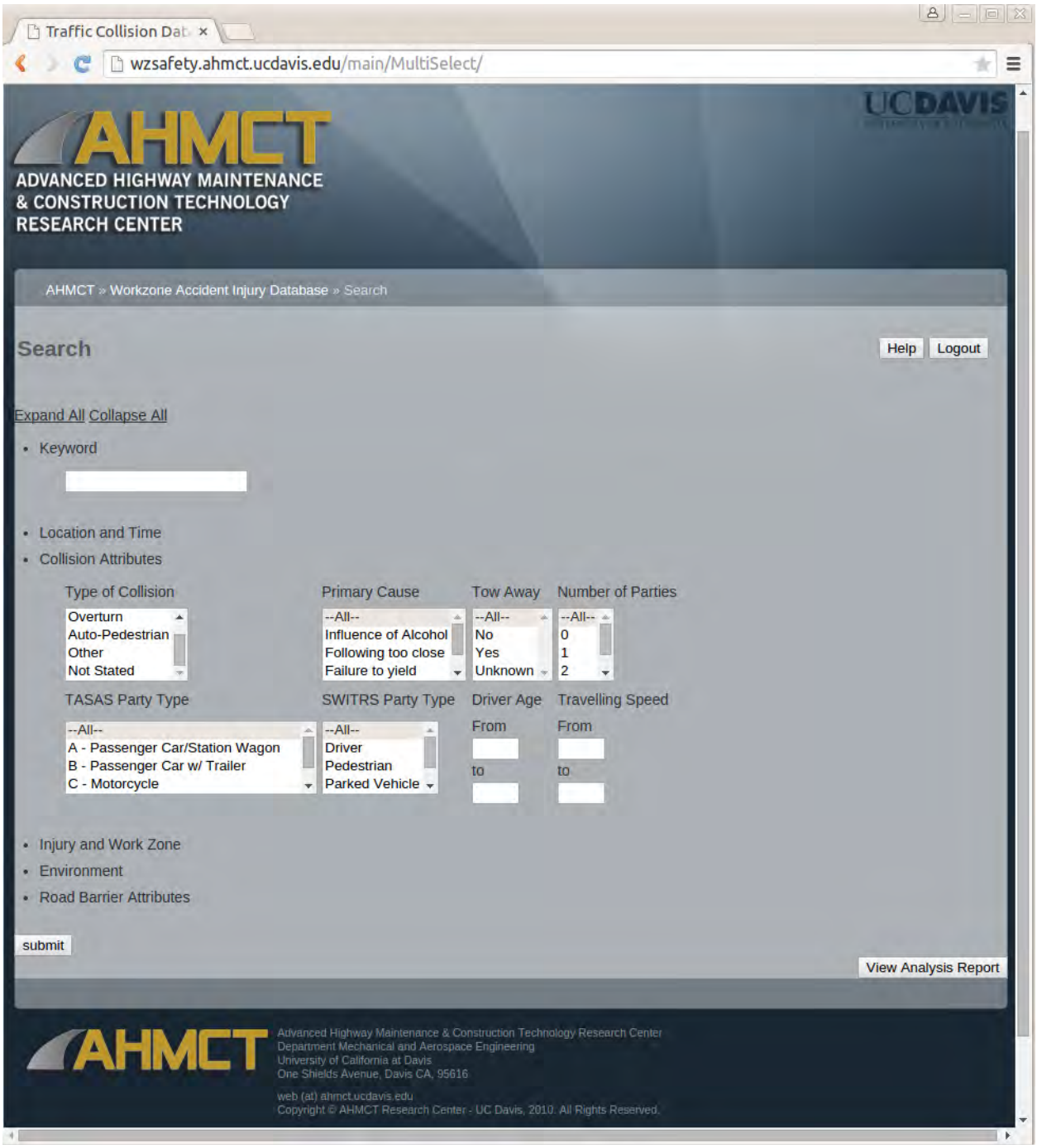


Figure 16. Searching By Categories.

Results Page

Once a search category is selected, then the user can click the *submit* button. The next page displays a table of all reports that match the search criteria. By clicking on each field of the top row of the table, the user can sort the table contents by that column. Also, the user can search for any keyword through the table using the *search box*.

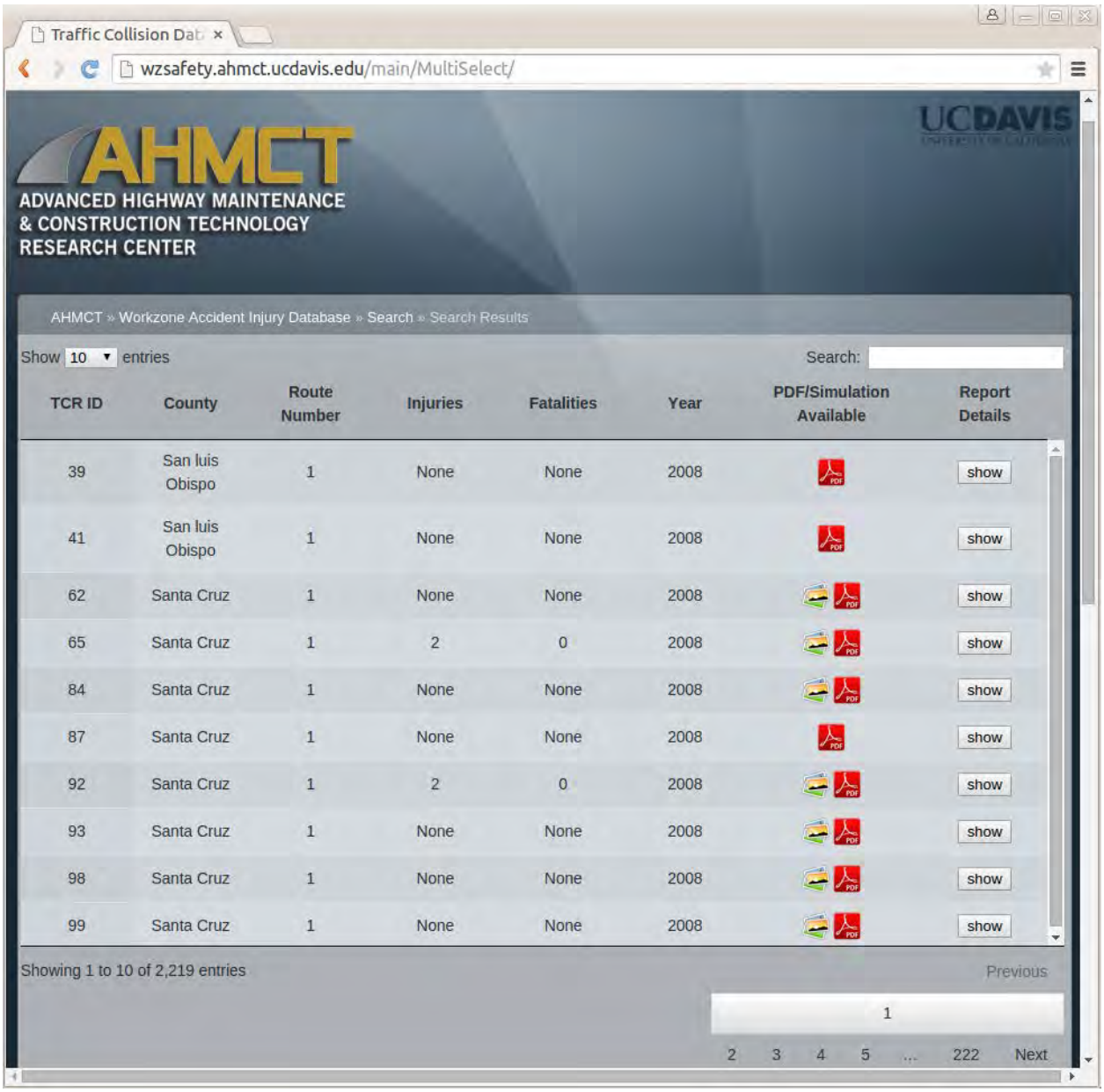


Figure 17. Reports Matching Search Criteria.

At the bottom of this page, a clickable map shows the location of all reports in the table (as shown in Figure 18). Clicking on each circle will show the reports within that area. This kind of display can be used, for example, to identify hot spots in a district where there has been more work zone accidents. The hot spot areas can then be further studied to determine the potential reasons for the higher distribution of accidents in work zones in such areas.

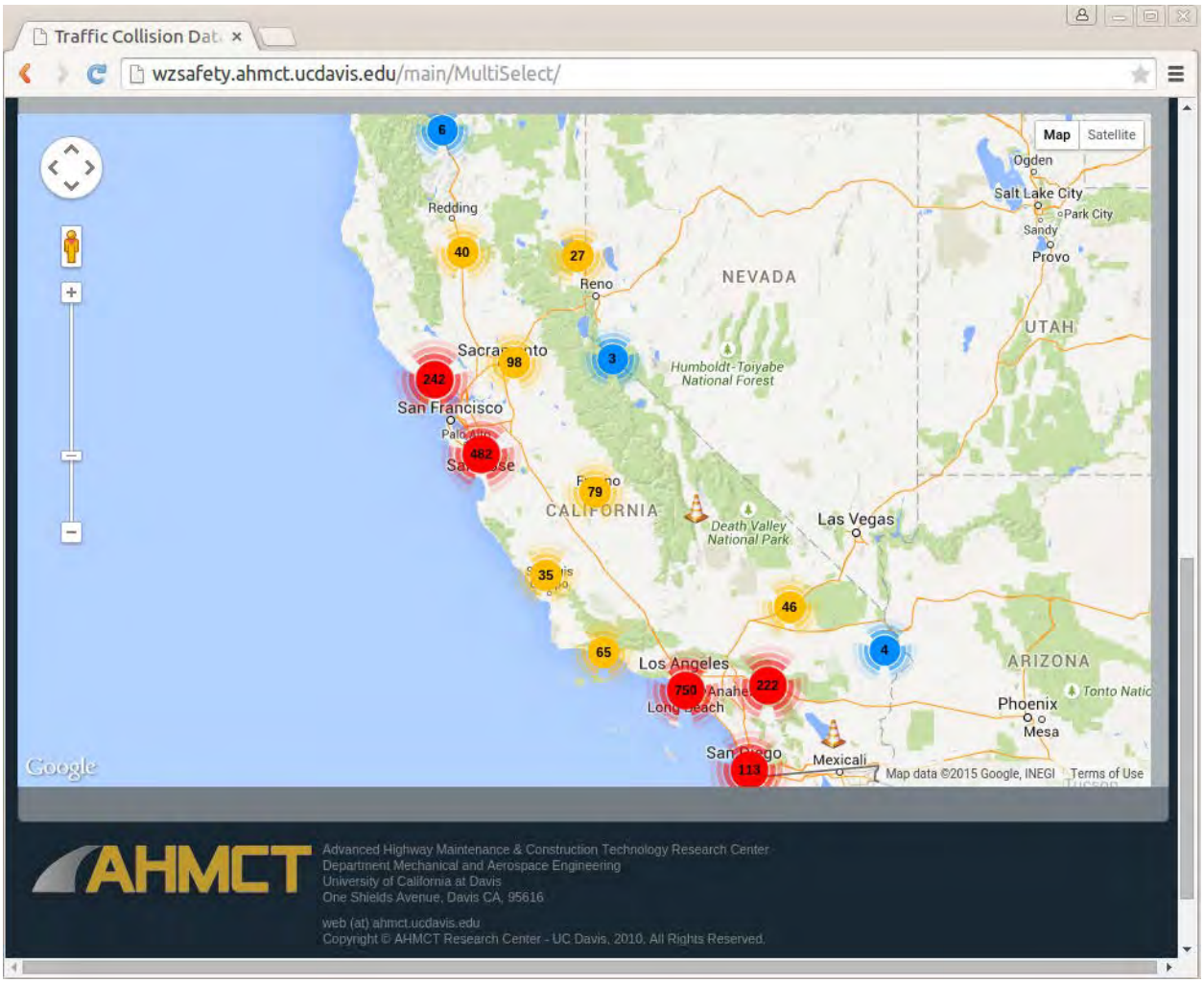


Figure 18. Mapping of All Searched Reports to the Accident Locations.

Displaying Report Details

In performing the analysis of the data in small sample of accidents, the accident description in the TCR was used to develop a simulation of the accident using the software package PC-CRASH. In such cases the PC-CRASH simulation was also added to the data base. In fact, in Figure 17, the second to the last column of the table of the search results has an indicator when a simulation exists for the accident in the data base. By clicking on the show button of the last column in the table of Figure 17, the user can open the accident details in a new page. In this page, as shown in Figure 19, a summary of accident details is provided on the left side and the location of the collision is shown on the map. The user can open the scanned pdf of the TCR by clicking on the *"view scanned report"* button.

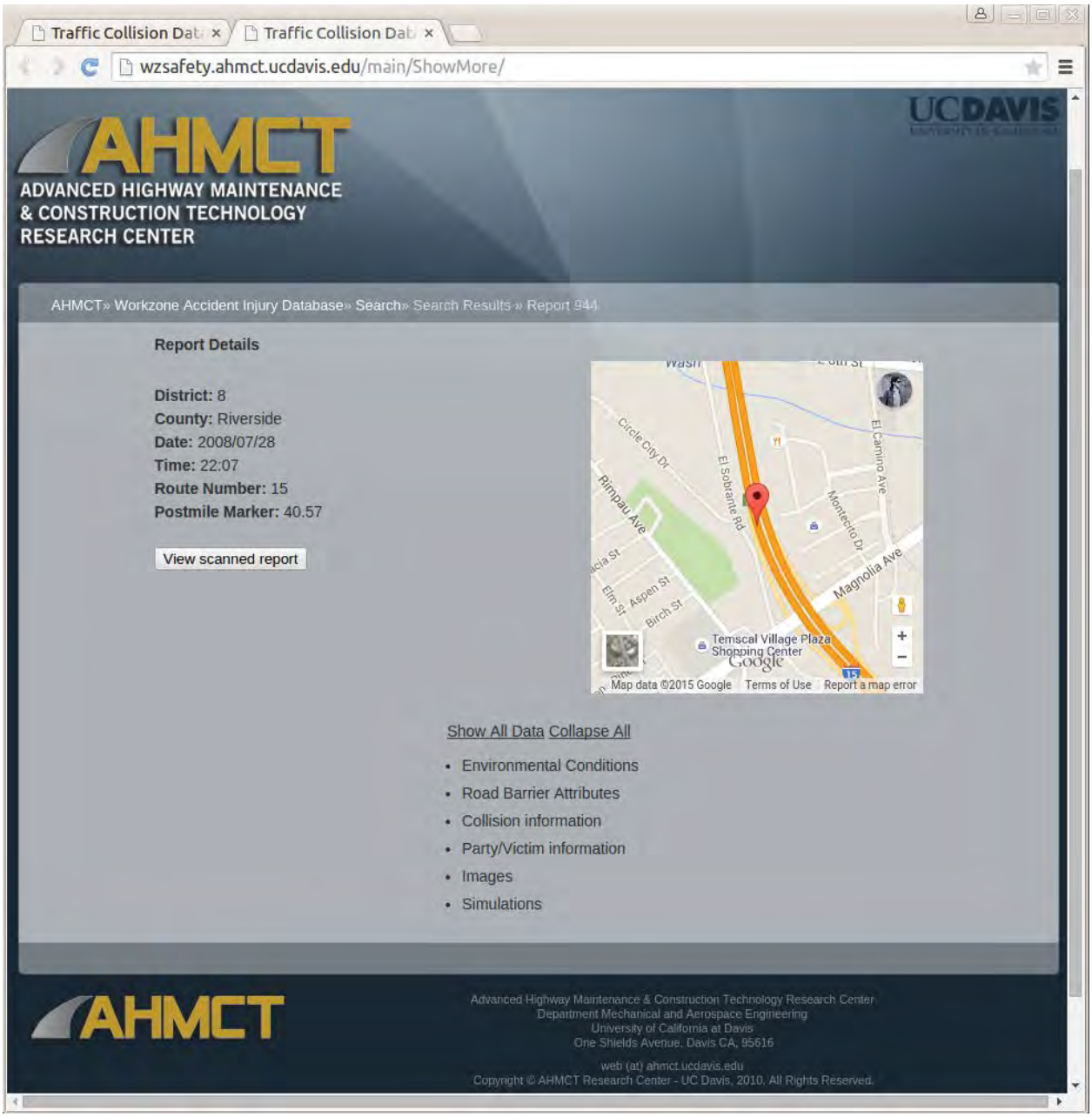


Figure 19. Accident Report Page.

Here, more accident details are also available to the user within these categories:

- Environmental conditions
- Road-Barrier Attributes
- Collision Information
- Party/Victim Information
- Images
    - Factual collision diagram from the report(if available)
- Simulations
    - PC crash simulation of the report(if available)

The user can view the details within each category by clicking on its name. Also, clicking on *"Show All Data"* button would display all details of the accident as coded. An example of the output from a PC-CRASH simulation of an accident within the data base is provided in Figure 20. The output would consists of a video type simulation of the accident depicted in the lower portion of Figure 20 as well as bird eye view of the collision in the upper portion of the same figure.

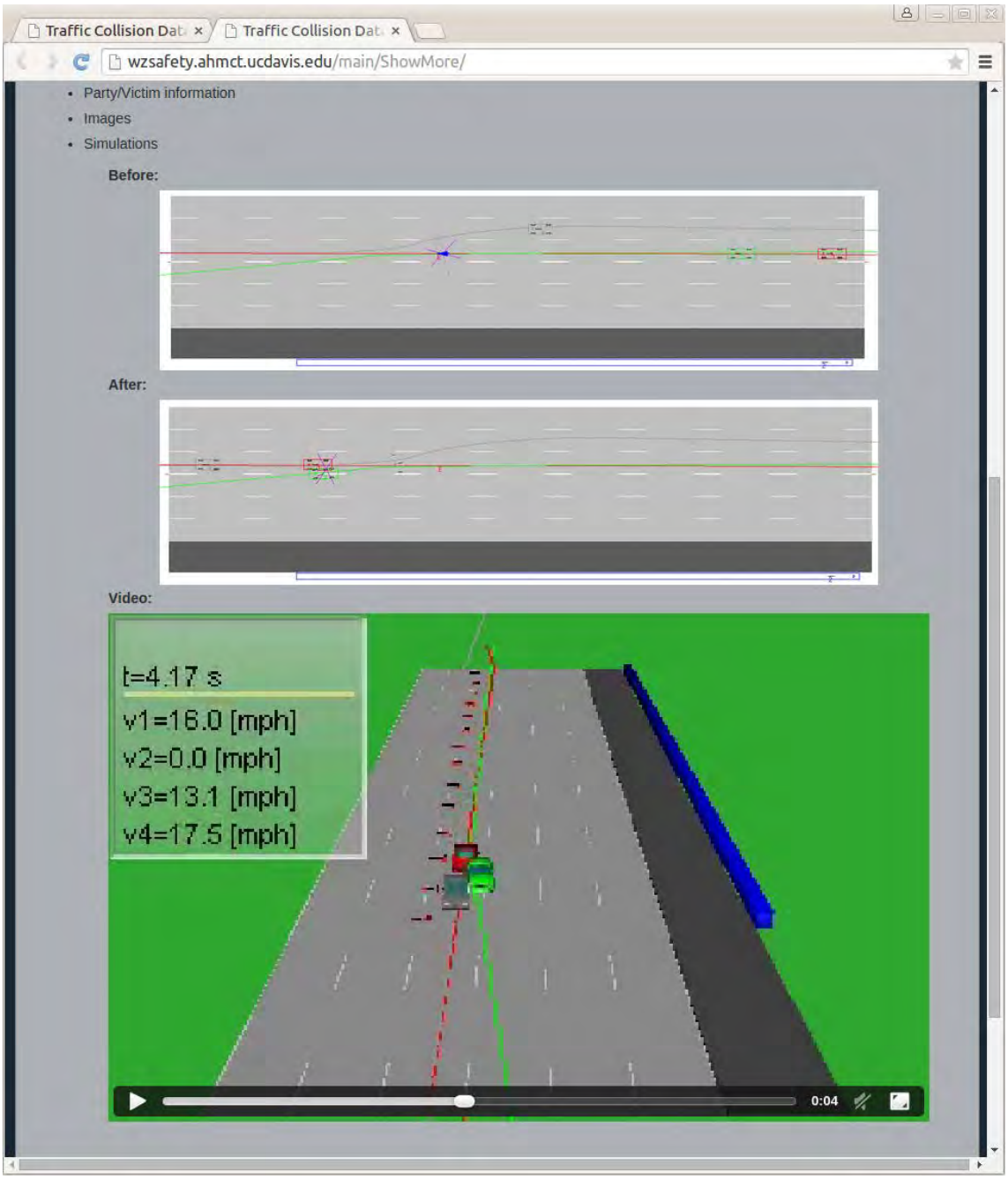


Figure 20. Output of a PC-CRASH Simulation.

DATA ANALYSIS

One significant analytical capability developed in this research study and coded into the decision support tool is the ability to calculate cost due to traffic collisions in work zones. The injury cost basis used in this particular analysis can be seen in Table 5 and is based on crash severity (DOT, 2015). Alternate cost models can be easily applied such as those which include non-medical economic impacts due to injury or those based on MAIS (Maximum Abbreviated Injury Scale) values of injured persons. To illustrate the cost analysis capabilities for this report however, only the model shown in Table 5 will be applied.

Table 5- Cost Model based on crash severity.

Crash Severity	Economic Cost
Fatality	$5,800,000
Serious Injury	$82048
Visible Injury	$23742
Complaint of Pain	$19494
Property Damage Only (PDO)	$10,439

Applying this cost model, the resulting total injury costs for all work zone accidents in California per year for the five years considered are calculated and summarized in Table 6. The values are broken down by year (2006 to 2010) and by Caltrans District. The values shown here include all injured and fatalities multiplied by the crash severity associated cost.

As can be seen from Table 6, there is a decreasing general trend from year 2006 (peaking at $547 million) down to $292 million in 2010. The total amount for all 5 years from work zone collisions is over $1.9 billion.

The total cost distribution based on crash severity can be seen in Figure 21. It is easily seen here that the cost of fatalities far outweigh the cost of all other crash severities combined using the cost model described earlier. Daytime collisions (Figure 22), the proportion amongst injury costs is very similar to those seen in the total averages (Figure 21). At nighttime (Figure 23), it can be seen that the proportion for fatality injuries increases slightly. The fatalities in both day and night time accidents account for approximately 75% of the collisions.

Table 6- Total Work Zone Collision Costs broken up by year and Caltrans District.

	2006	2007	2008	2009	2010	Total (million)
District 1	$12,248,922	$30,441,589	$13,172,586	$1,159,640	$1,164,240	$58
District 2	$13,236,764	$7,394,284	$7,214,313	$1,349,278	$1,692,861	$31
District 3	$19,879,369	$11,065,107	$31,342,978	$27,264,298	$50,046,310	$140
District 4	$34,136,944	$54,388,895	$54,236,783	$68,477,125	$30,834,780	$242
District 5	$5,108,450	$22,643,147	$17,629,245	$4,135,722	$8,918,007	$58
District 6	$27,073,443	$18,167,639	$9,537,151	$49,782,197	$9,143,546	$114
District 7	$99,615,822	$45,915,519	$96,895,023	$60,432,125	$89,617,434	$392
District 8	$153,862,299	$95,174,007	$31,083,888	$51,940,493	$52,486,316	$385
District 9	$6,077,158	$20,878	$6,296,585	$662,241	$172,848	$13
District 10	$28,923,395	$37,079,908	$36,587,507	$4,915,875	$8,170,412	$116
District 11	$87,472,479	$20,877,010	$29,899,499	$59,641,170	$30,271,865	$228
District 12	$59,438,775	$28,280,003	$17,192,920	$16,772,022	$9,623,974	$131
Total (million)	$547	$371	$351	$347	$292	**$1,908**

Total Costs Average per year ($1908 million / 5 years) = $382 million per year

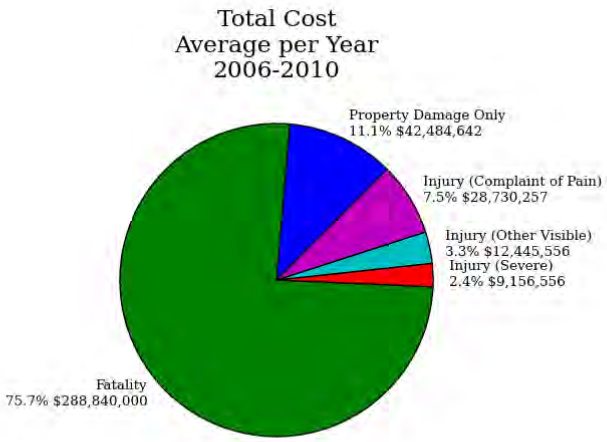


Figure 21- Distribution of work zone collision costs for the time period from 2006 through 2010.

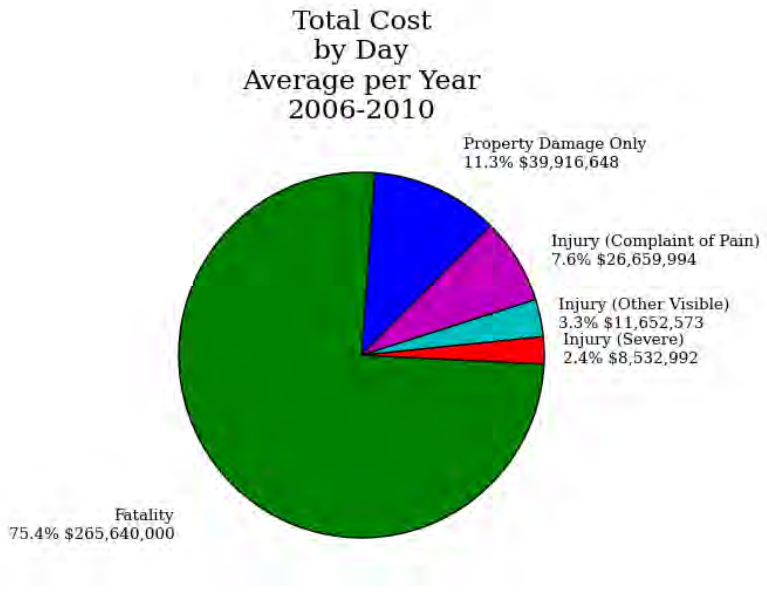


Figure 22. Total work zone collision cost for Daylight only collisions.

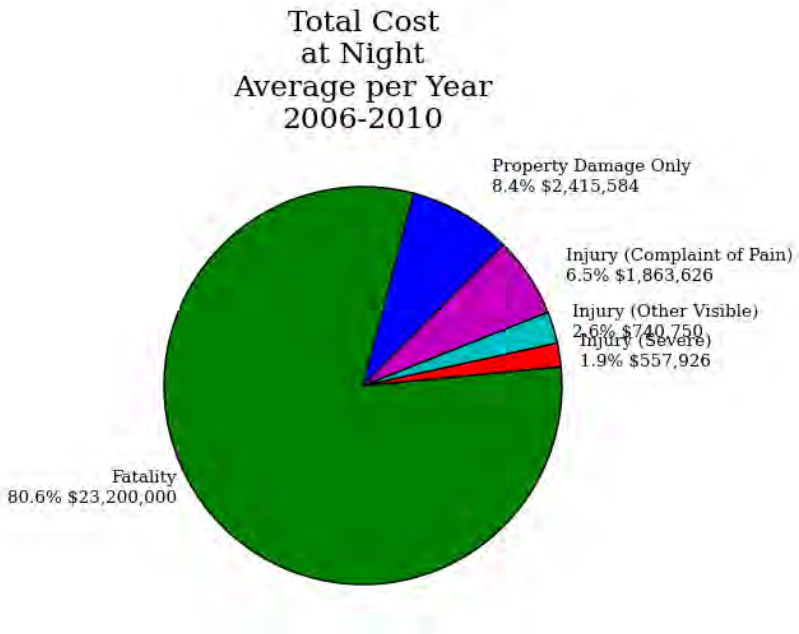


Figure 23. Total work zone collision cost for Nighttime only collisions.

When comparing total numbers of collisions (averaged over the years 2006-2010) it can be seen that property damage only is the most common and occurs 80% of the time (Figure 24). When considering the average collision counts per year for Primary Cause factors (Figure 25) it can be seen that" speeding/following too closely" is responsible for 50% of collisions with "Improper Driving" being a close second at 37%. For collisions occurring at night (Figure 26) there are some differences but the general trends stay the same.

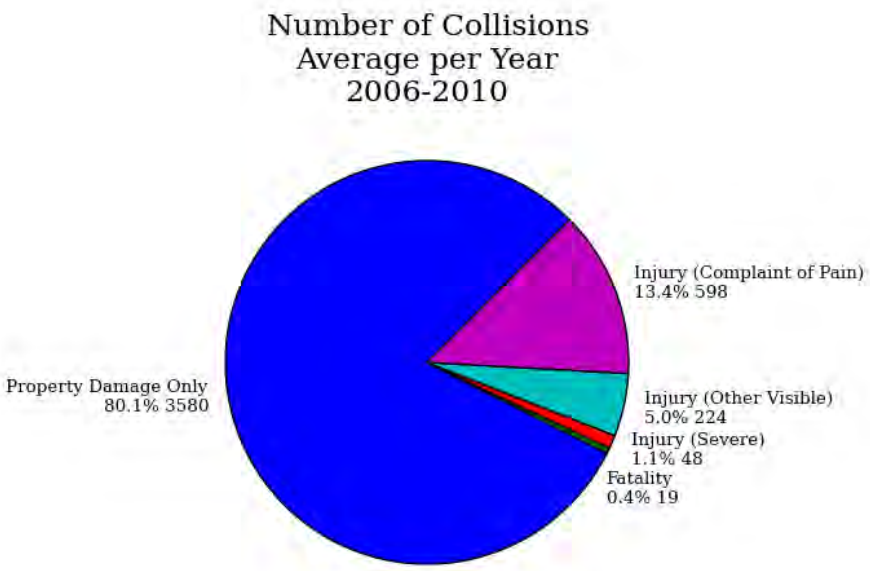


Figure 24- Number of collisions averaged per year shown as a percentage of totals.

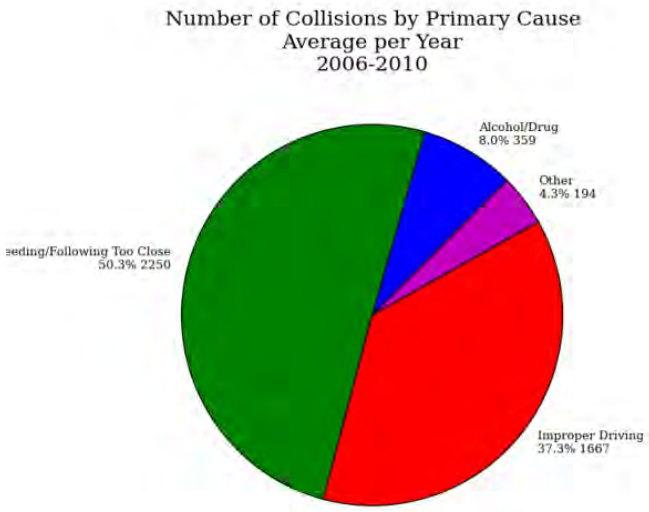


Figure 25- Pie chart of collision frequency per primary cause.

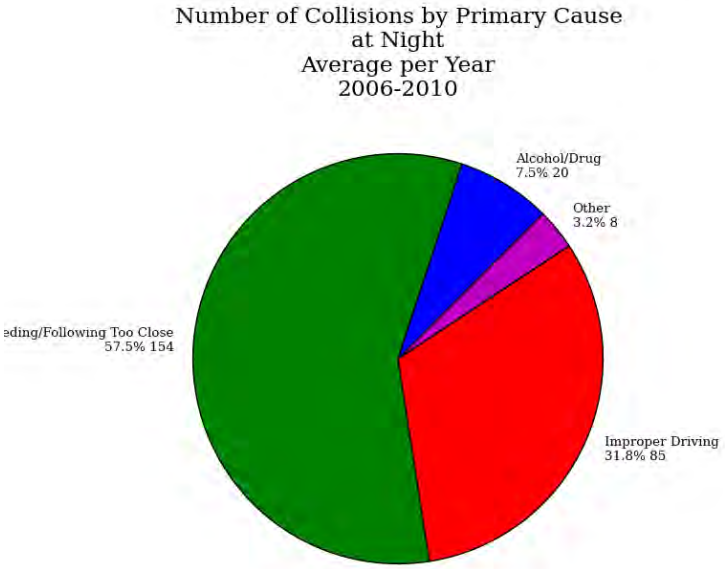


Figure 26. Pie chart of collision frequency per primary cause.

BIBLIOGRAPHY

DOT. (2015). *The Economic and Societal Impact Of Motor Vehicle Crashes, 2010 (Revised).* Washington, DC: Department of Transportation.

State of California Department of Transportation. (2014). Chapter 3: Accident and Roadway Records. In C. T. Operations, *Traffic Manual.*

APPENDIX A

Automatic Traffic Collision Report Optical Text Recognition Analyzer

Due to the high volume of collected traffic collision reports, it was decided to develop automated or semi-automated ways that at least some of the information can be extracted and codified. Optical Character Recognition (OCR) software was therefore developed to aid the process. This appendix describes the OCR software developed as a byproduct of this research. It should be pointed out that this software development was not part of this research study. It was only undertaken to make sure that the research completion deadline can be met and all TCRs are codified by automating some of the process.

Open-Source Dependency

For ease of future potential portability of the optical character recognition software to Caltrans, it was decided that open-sourced software and library packages would be utilized. Specifically, the following packages are used:

- Tesseract
- PyEnchant
- Spelling Checker
- PyPDFOCR
- Google's Geocode Database

Each of these is described in more detail in the subsequent section of this appendix.

Tesseract

Tesseract is an open-source Optical Character Reader/Recognition engine for Windows, OSX, and Linux Operating Systems. The software is released under the Apache License 2.0. Tesseract is considered as one of the most accurate open-source OCR software available to the public.

Tesseract OCR engine was originally developed by Hewlett Packard (HP) from 1985 – 1994. Tesseract was one of the top 3 OCR engines in 1995 and tested by University of Nevada, Las Vegas (UNLV). Between the periods of 1985-1998, Tesseract OCR engine was written in C programming language and some portions were translated to C++ programming language in 1998. However, there was little amount of work done between periods of 1995 – 2006. Tesseract released as open-source software by HP and UNLV in 2005. Later in 2006, Google

started to sponsor the development of Tesseract. Since then, the development process and accuracy of Tesseract OCR engine have improved extensively. Tesseract OCR engine combines with the Leptonica Image Processing Library, which allows the engine to decompress and read a variety of compressed/-uncompressed images over 60 programming languages.

Detailed information on Tesseract:

- Latest Release- Please see Tesseract Release Notes (*https://code.google.com/p/tesseract-ocr/wiki/ReleaseNotes*) for details.
- Download: For the latest download, please see Tesseract Source Download *(https://drive.google.com/a/ucdavis.edu/folderview?id=0B7l10Bj_LprhQnpSRkpGMGV2eE0&usp=sharing)* for details.
- For more Information: go to *https://code.google.com/p/tesseract-ocr/*

PyEnchant

PyEnchant is an open-source spellchecking library based on Enchant (see *http://www.abisource.com/projects/enchant/*). PyEnchant combines all the functionality of the underlying Enchant library with the flexibility of Python and a nice "Pythonic" object-oriented interface.

Detailed information on PyEnchant:

- Download: For the latest download, please see *(http://pythonhosted.org/pyenchant/download.html)* for details.
- For more Information: go to https://code.google.com/p/tesseract-ocr/ PyEnchant Homepage (http://pythonhosted.org/pyenchant/)

Spelling Corrector

Spelling Corrector is a simple open-source English language spelling corrector implemented by Peter Norvig. The corrector provides an overview and simple spelling correction concept that most web-based search engines use. The idea is to be able to provide spelling suggestion when users entered misspell words. The theory of how this script works is the same as Google's "Did you mean". Spelling Corrector is able to provide spelling suggestion up to 10 words per second.

By combining the simple machine learning theory, the spelling corrector takes in training data (text of some sort) and counts the frequency of each word appears in the text. By splitting, deleting, transposing, replacing, and inserting alphabet characters to the words that provide, spelling corrector will be able to generate a list of possible "words" (words can be non-English words) that the spelling is closely related to the original word. Finally, spelling corrector applies

some statistical analysis, such as maximum likelihood to the filter words that is not English and less likely to be used.

For more information see How to write a Spelling Corrector *(http://norvig.com/spell-correct.html).*

PyPDFOCR

PyPDFOCR is an open-source library that allows users to convert image-based Portable Document Format (PDF) files into OCR-ed (searchable image base) PDF files.

PyPDFOCR firstly converts the PDF document into multiple pages Tiff Image, and then converts the Tagged Image File Format (TIFF, uncompressed image, see TIFF for details) into Joint Photographic Experts Group Image (JPEG, compressed image, see JPEG for details) by Ghostscript. Next, PyPDFOCR performs an OCR operation with Tesseract on each individual image to extract text and store in HOCR (html-based text file). Finally, PyPDFOCR rebuilds the PDF document by the HOCR files.

Detailed information on Tesseract:

- Download: For the latest download, please see *(https://github.com/virantha/pypdfocl)* for details.
- For more Information: go to https://code.google.com/p/tesseract-ocr/ PyPDFOCR GitHub Repository (*https://github.com/virantha/pypdfocr).*

Google Services

Google's Reverse Geocoding is a process to convert latitude and longitude location into a human readable address format. Google's Reverse Geocoding Database allows you to find the name of the location.

For more information see Google Geocode (*https://developers.google.com/maps/documentation/geocoding/*) for details.

Direction API

Google's Direction Services is a service to calculate direction between locations over a HTTP request. Multiple direction calculation modes are available: transit, driving, walking or cycling. Google's Direction Services is designed for calculating direction and distance between two-fixed static locations; it is NOT designed for real time data calculation.

For more information see Google Direction (*https://developers.google.com/maps/documentation/directions/) for* details.

Tesseract Training Tools

Tesseract was originally designed to recognize English text only. Efforts have been made to modify the engine and its training system to make them able to deal with other languages and ASCII (American Standard Code for Information Interchange) characters (Tesseract Training Tools, https://code.google.com/p/tesseract-ocr/wiki/TrainingTesseract3). Note: Only Tesseract3 is trainable.

Training Procedure

Tesseract training method is fully depended on box files, where it contains coordinates of characters in the image. The training guide from Tesseract website can be handy. However, AHMCT had put together python scripts to speed up the process.

Step 1: create training image of the format you want.
Note: that training image work the best in TIF and TIFF format.

Step 2: create box files by executing the following command on each image:
tesseract <imag path> <training name> **batch.nochop makebox**
OR
Create box files by executing the AHMCT python make box script:
python tesseract_makebox.py <directory_path>
Where **directory_path** is the path to the folder containing all the image files.

Step 3: manually edit box files to correct whatever mistake Tesseract made.
See next section for guide on editing box files.

Step 4: When all editing is complete, execute the following command
python tesseract_training.py <directory> <language name> <font name> <italic> <bold> <fixed> <serif> <fraktur>
Where italic, bold, fixed, serif, and fraktur are 0/1 indicating true or false.
Example:
python tesseract_training.py ./**wzi wzi tcr 0 1 0 0 0**
Above command train language as wzi, font name tcr, and is bold.

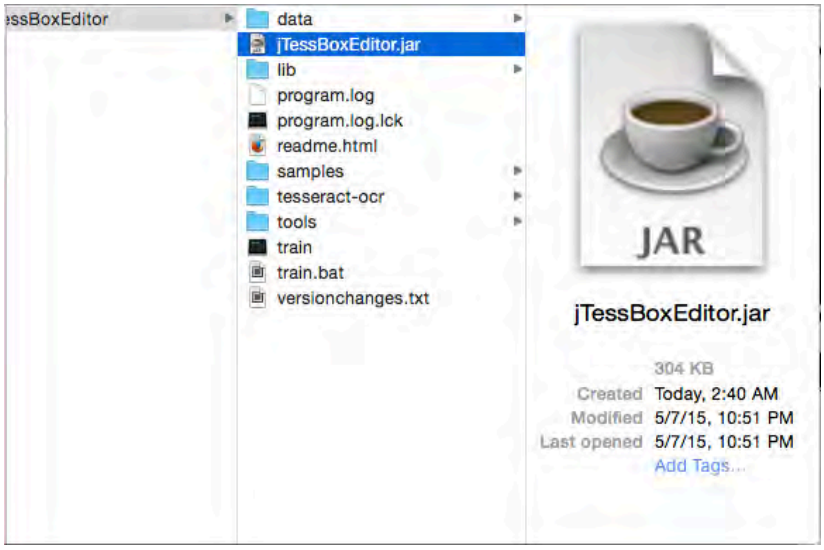
Step 5: copy **<language>**.trainddata to tessdata located at:

/usr/local/share/tessdata

Editing Box Files

There are several useful tools for editing Tesseract box files. In this section we will be using jTessBoxEditor.

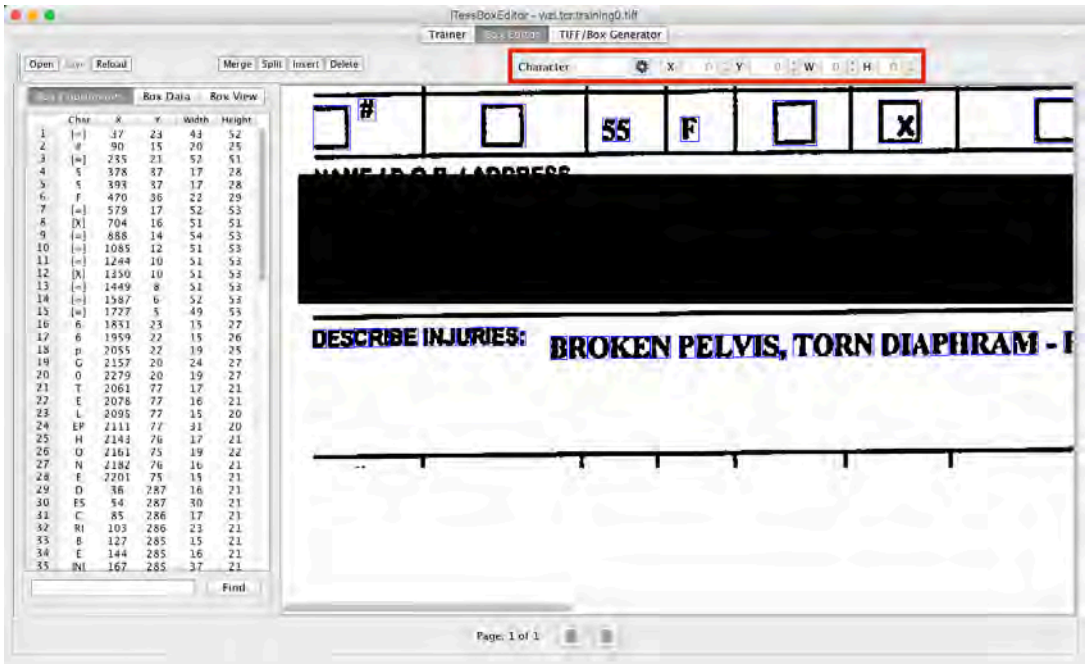
Open jTessBoxEditor by double click on **jTessBoxEditor.jar**



On the upper tab, navigate to **Box Editor**

Open an image file by clicking open on the upper left

Adjust the position and correct representation of the blue box according to the image by using the setting manual on top of the image (highlighted in red).



Traffic Collision Report Processing Program

The main program governing the execution of the software and controlling the conversion process is referred to as the "Traffic Collision Report Processing Program" (TCRPP). The overall operation of this Python program is described in the following section of this document.

Initial Processing of the Traffic Collision Report

The input format of each collision report is in PDF format. In order to "read" the files virtually, each page of the report much be converted into a TIFF image. Throughout the program entire pages or portions of a page are input into "Tesseract" to extract any text contained in the image.

Tesseract then returns the memory address of where the OCR-ed text is stored on the system memory to python. Processing of the output must be done in order for it to be recognizable and manipulated as a character string.

Text correction

In order to provide accurate information from the extracted "text", multiple layers of spelling check are being performed before analysis takes place. TCRPP filters all texts or symbols that do not make sense within the sentence; such as "State of California.''sd;';';" will be corrected to "State of California". This functionality is capable of continually improving as more and more

reports are converted. PyEnchant is used here to verify valid English words along with "Spelling Corrector" to identify and accept abbreviations, acronyms, etc. All text correction processes are done in Text Pattern Analyzer, described further on.

Spelling Corrector

Text rejected by PyEnchant will be passed onto the Spelling Corrector program. Please recall the Spelling Corrector is continually "trained" and maintains its own "dictionary" geared specifically for the TCRPP program. Words such as "State of California" might be recognized as "tat of California" or "tat of Calif0rni4", etc. Spelling Corrector take the statistical measurement on what characters from a word is most frequently missing from Tesseract, and how likely does it occur on reports. Spelling Corrector will correct the word according to the statistic.

Miss-OCR Analyzer

Miss-OCR Analyzer is a package developed by AHMCT to correct commonly miss-OCRed text from the report image, such as "A" was recognized as "4". However, this operation is done within Collision Information Extraction, since the correction must be done according to the information being extracted.

Collision Information Extraction

Once text has been extracted, filtered and verified, a set of operations are then called to extract the desired information from the original text.

Text Pattern Analyzer

Text Pattern Analyzer is designed for TCRPP to catch any non-valid words line-by-line, word-by-word from Tesseract. This step of the program analyzes individual words and determines if they are numbers, accepted words or "names".

Page Type Analyzer

Page Type Analyzer recognizes the Page Header on each page (see Figure 27) and then discerns the form type within the Traffic Collision Report. Header types are: The following list gives the possible Header Type along with the name of the function-based program:

1. Traffic Collision Report – Collision Information Analyzer, Party Information Analyzer, Driver Information Analyzer, Towed Information Analyzer, Vehicle Damage Analyzer, Safety Equipment Analyzer.
2. Traffic Collision Coding – Currently Not Supported.
3. Diagrams – No Analysis needed.
4. Injured/Witnesses/Passengers – victims Information Analyzer.
5. Narrative/Supplemental – No Analysis needed.
6. Truck/Bus – Current Not Supported.
7. Physical – Current Not Supported.

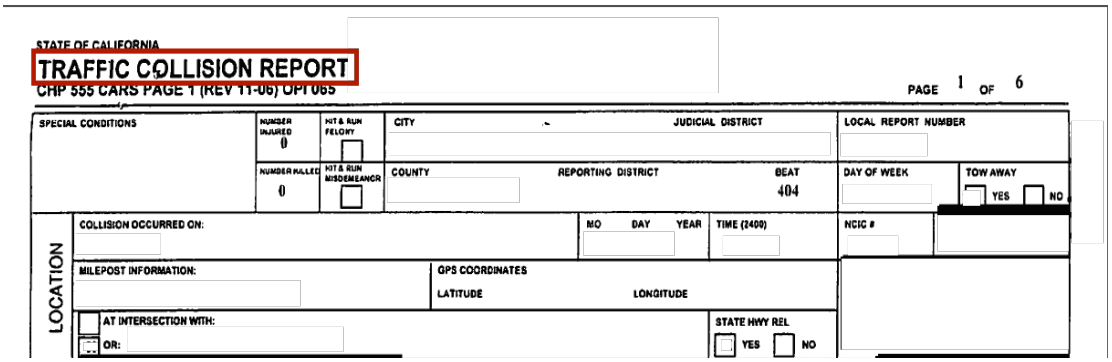

STATE OF CALIFORNIA
TRAFFIC COLLISION REPORT
CHP 555 CARS PAGE 1 (REV 11-06) OPI 065
PAGE 1 OF 6
SPECIAL CONDITIONS
NUMBER INJURED 0
HIT & RUN FELONY
CITY
JUDICIAL DISTRICT
LOCAL REPORT NUMBER
NUMBER KILLED 0
HIT & RUN MISDEMEANOR
COUNTY
REPORTING DISTRICT
BEAT 404
DAY OF WEEK
TOW AWAY YES NO
LOCATION
COLLISION OCCURRED ON:
MO DAY YEAR
TIME (2400)
NCIC #
MILEPOST INFORMATION:
GPS COORDINATES
LATITUDE
LONGITUDE
AT INTERSECTION WITH:
OR:
STATE HWY REL YES NO

Figure 27. Extracting the collision data from the first page of TCR.

<u>Collision Information Analyzer</u>

Collision Information Analyzer is a program developed in this research study that uses as its input the ASCII data from OCR. This analyzer locates and extracts basic collision information from the upper section of the first page of the traffic collision report. Information extracted includes: county, highway number, pre or post mile designation, post-mile marker, and landmark description of collision. These areas are highlighted in Figure 28. "Reading" this basic information is not trivial and three basic operations are performed to achieve this as follows:

1. First Operation

The first operation of the Collision Information Analyzer executes ONLY if either latitude or longitude by the GPS Location Analyzer or Day of Week was located on the report. Collision Information will be extracted under "Collision Occurred On", such as county, highway number, post-pre, and post-mile marker. By locating the county, TCRPP is able to map the county into its associate district.

2. Second Operation

If no result returns from the first operation, TCRPP will perform the second operational loop of collision information extraction. This operation simply locates known keywords of acronyms. Currently, the second operation stores only highway numbers and counties as reference.

3. Third Operation

The third operation will run only if the latitude and longitude are provided in the "Location" section of the Traffic Collision Report. Google's Direction API is applied along with a program that AHMCT developed mapping all Caltrans highways into longitudinal-latitudinal coordinates.

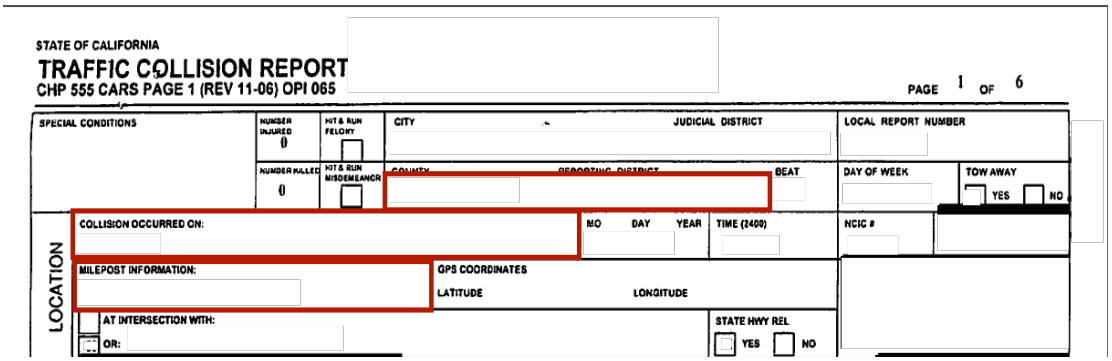


Figure 28- Portions highlighted from Page 1 of a TCR where basic collision information is located.

Date Time Analyzer

Date Time Analyzer locates the date and time of the collision from the top section of the report; see the example in Figure 29. Date and time is then verified by python built-in date-time library after being located. Date and time are located on all subsequent report pages (see Figure 30).

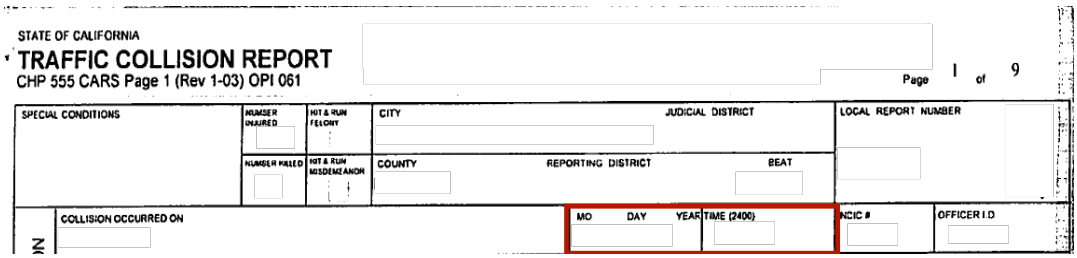
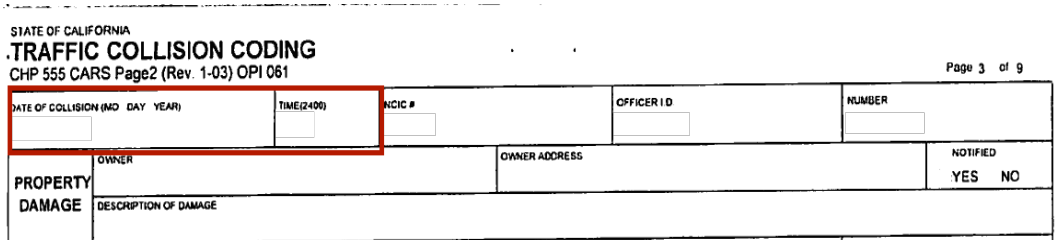


Figure 29. Highlighted region in a TCR indicates where date and time are located


STATE OF CALIFORNIA
.TRAFFIC COLLISION CODING
CHP 555 CARS Page2 (Rev. 1-03) OPI 061
Page 3 of 9

DATE OF COLLISION (MO DAY YEAR)	TIME(2400)	NCIC #	OFFICER I.D.	NUMBER

PROPERTY DAMAGE	OWNER	OWNER ADDRESS	NOTIFIED
	DESCRIPTION OF DAMAGE		YES NO

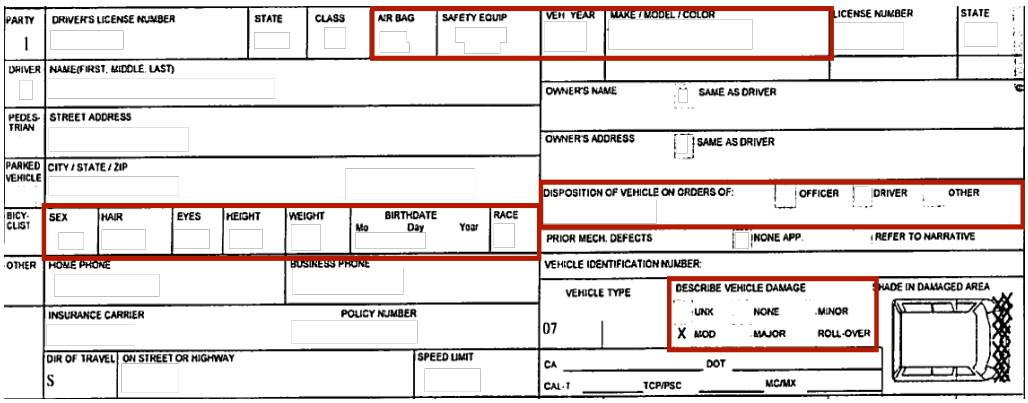
Figure 30- Highlighted region in a TCR indicates where date and time are located on "all the other pages".

Party Information Analyzer

Party Information Analyzer performs analysis on second to fourth sections of the first page of the Traffic Collision Report; see Figure 31 for details.

Party Information Analyzer includes three layers of information extraction/recognition, including Vehicle Information Analyzer, Driver Information Analyzer, Vehicle Damage Information Analyzer, and Towed Information Analyzer.

Information extracted includes: Air Bag, Safety Equipment, Vehicle Year, Vehicle Make/Model/Color, Driver Sex, Driver Hair Color, Driver Eyes Color, Driver Height, Driver Weight, Driver Birthday, Driver Race, Towed/Driven Away, and Vehicle Damage


PARTY 1 | DRIVER'S LICENSE NUMBER | STATE | CLASS | AIR BAG | SAFETY EQUIP | VEH YEAR | MAKE / MODEL / COLOR | LICENSE NUMBER | STATE

DRIVER | NAME(FIRST, MIDDLE, LAST)

PEDES-TRIAN | STREET ADDRESS

PARKED VEHICLE | CITY / STATE / ZIP

BICY-CLIST | SEX | HAIR | EYES | HEIGHT | WEIGHT | BIRTHDATE Mo Day Year | RACE

OTHER | HOME PHONE | BUSINESS PHONE

INSURANCE CARRIER | POLICY NUMBER

DIR OF TRAVEL S | ON STREET OR HIGHWAY | SPEED LIMIT

OWNER'S NAME | SAME AS DRIVER

OWNER'S ADDRESS | SAME AS DRIVER

DISPOSITION OF VEHICLE ON ORDERS OF: OFFICER DRIVER OTHER

PRIOR MECH. DEFECTS | NONE APP. | REFER TO NARRATIVE

VEHICLE IDENTIFICATION NUMBER:

VEHICLE TYPE 07

DESCRIBE VEHICLE DAMAGE: UNK NONE MINOR X MOD MAJOR ROLL-OVER

SHADE IN DAMAGED AREA

CA DOT CAL-T TCP/PSC MC/MX

Figure 31- First page of a TCR with party information highlighted.

Vehicle Information Analyzer

Vehicle Information Analyzer locates and extracts information related to the vehicle(s) involved in the collision. Information extracted includes: Air Bag, Safety Equipment, Vehicle Year, and Vehicle Make/Model/Color. Note that each vehicle is associated with a party number. See Figure 32 for details.

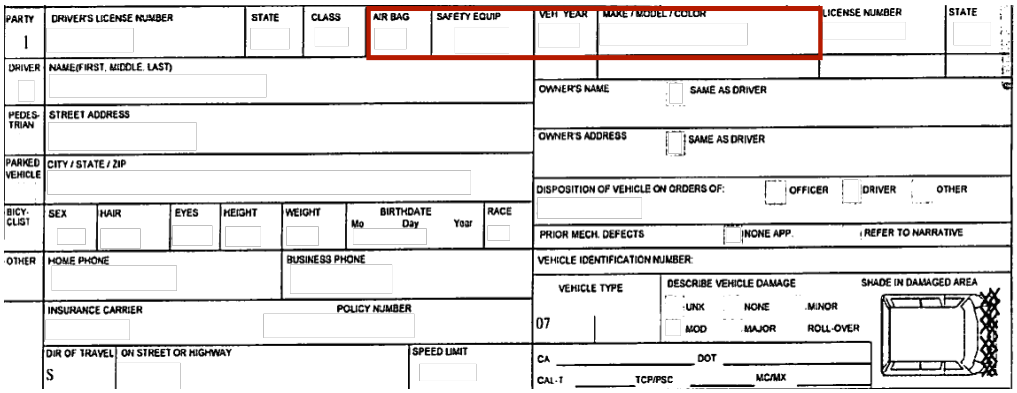


Figure 32- Vehicle information is highlighted on the first page of the Traffic Collision Report. Note the vehicle information is always associated with a party number.

Driver Information Analyzer

Driver Information Analyzer locates and extraction/recognition information related to the driver of the vehicle involved in the collision associated to the party. **Information extracted includes:** Driver Sex, Driver Hair Color, Driver Eyes Color, Driver Height, Driver Weight, Driver Birthdate, and Driver Race (Figure 33).

Example of Driver Information Location on a TCR is provided in Figure 33.

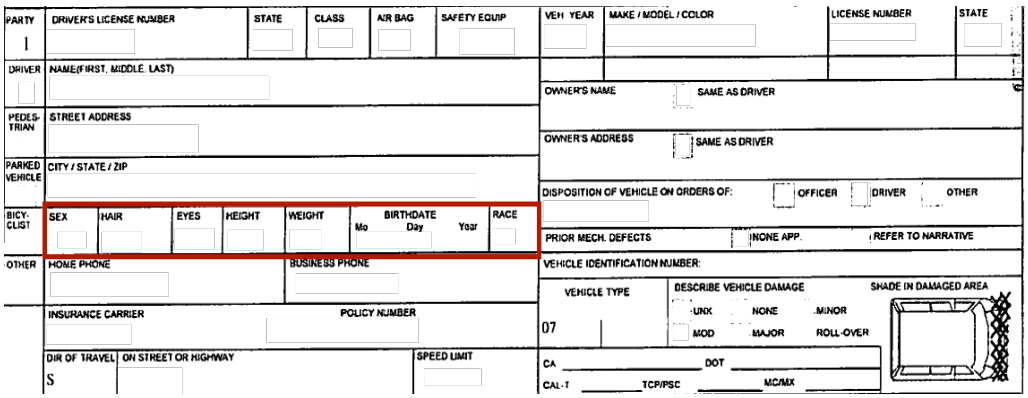


Figure 33- The region where the driver information is highlighted is illustrated here.

Vehicle Damage Information Analyzer

Vehicle Damage Information Analyzer locates and extracts vehicle damage information related to the vehicle involved in the collision associated to the involved party.

Information extracted includes: Vehicle Damage
Vehicle Damage: UNK, None, Minor, MOD, Major, Roll-Over

Example Vehicle Damage Analyzer: Figure 34

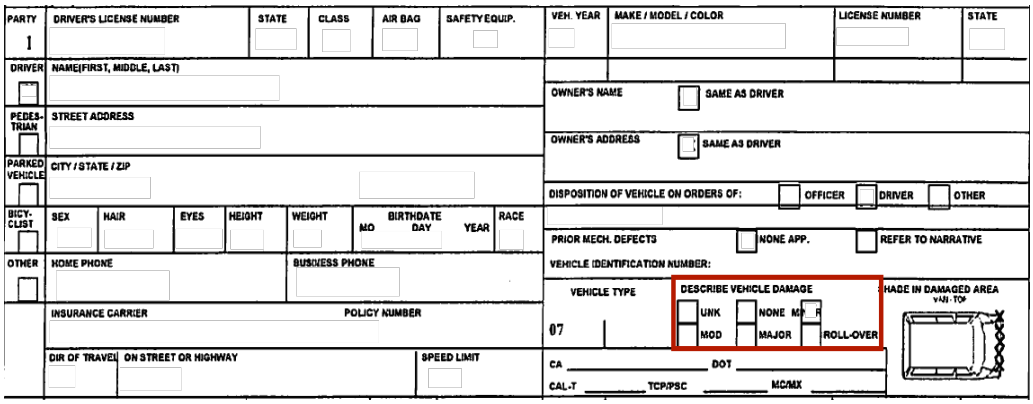


Figure 34- Region on the collision report where vehicle damage is designated.

Towed Information Analyzer

Towed Information Analyzer locates and extracts information on the vehicle involved in the collision which was towed away or driven away. Information extracted includes: Towed/Driven Away (Figure 35).

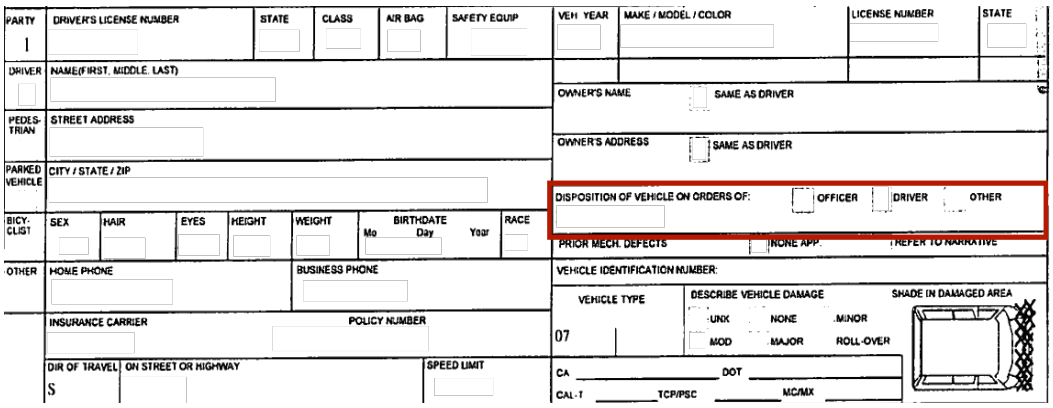
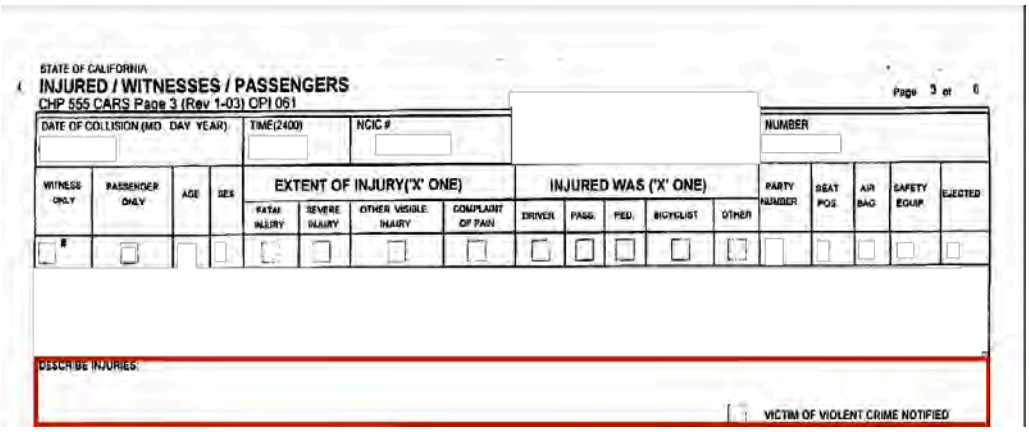


Figure 35- Towed or driven away for a particular vehicle is highlighted here.

Victim Information Analyzer

When the Page Type Analyzer identifies a page of the Traffic Collision Report that has the title of Injured/Witnesses/Passengers page (see Figure 36), TCRPP will perform victim information analysis on that page. Victim Information Analyzer locates and extracts the injuries information of victim(s) involved in the collision (see Figure 36). Information extracted includes: Describe Injuries.


STATE OF CALIFORNIA
INJURED / WITNESSES / PASSENGERS
CHP 555 CARS Page 3 (Rev 1-03) OPI 061

DATE OF COLLISION (MO. DAY YEAR)	TIME(2400)	NCIC #	NUMBER

WITNESS ONLY	PASSENGER ONLY	AGE	SEX	EXTENT OF INJURY('X' ONE)				INJURED WAS ('X' ONE)					PARTY NUMBER	SEAT POS.	AIR BAG	SAFETY EQUIP	EJECTED
				FATAL INJURY	SEVERE INJURY	OTHER VISIBLE INJURY	COMPLAINT OF PAIN	DRIVER	PASS.	PED.	BICYCLIST	OTHER					

DESCRIBE INJURIES:

VICTIM OF VIOLENT CRIME NOTIFIED

Figure 36- Highlighted area shows where victim injury is described in detail.

Installation/Configuration

System Requirement

Minimum Requirement

Operating System: Ubuntu 12.04 LTS
Graphic Card: N/A
Storage: 80GB
System Memory: see Chapter 10.1.3

Recommended Requirement

Operating System: Ubuntu 14.04LTS
Graphic Card: N/A
Storage: 160GB
System Memory: see Chapter 10.1.3

System Memory

The amount of system memory required for ATCRA is depending on the processes you execute on the machine. Please see the chart below.

Number of Processes	Amount of Memory Required (GB)
1	6
2	10
3	14
4	16

Installation

To install Automatic Traffic Collision Report Analyzer, execute the following command in Terminal:

tar –zxvf TCRA-<version>**.tar.gz**
cd TCRA-<version>
./configure

Execution

Single Report Execution

To perform information extraction/recognition on a single PDF report, execute the following command in Terminal:

Python traffic_collision_report_analyser.py [option] <pdf file path>

Multiple Reports Execution

To perform information extraction/recognition on multiple PDF reports or directory of reports, execute the following command in Terminal:

Python OCR_services.py [option] <directory/pdf file path>

Execution Option

This section will descript all execution option for ATCRA

- **-r** - ramdisk name, this option will take place only if ramdisk mode is enabled.
- **-e** – enable ramdisk mode, by default ramdisk mode is disabled.
- **-y** – report year, manually provide the year of report(s).

Known Bugs

- **Tesseract crashes**

Tesseract will sometime cause Automatic Traffic Collision Reports Analyzer (ATCRA) to crash due to the undefined memory buffer location. This crash occurs at random time, depending on the amount of memory the ATCRA requires per execution. This problem does not occur when running a single report.

The primary cause for this problem is that the Tesseract Application Programming Interface (API) was not designed for multiple OCR operation as a single process/program on the Operating System. However, during the initialization/startup of each process/program, operating system will allocate limited amount of memory for it. Since each Traffic collision report consists of different amount of page length, one that allocates for the current report might not be enough for the next one. Tesseract and ATCRA are both designed to overwrite the existing memory buffer instead of allocating new buffer for individual image. Therefore, if the amount of required memory is larger than the amount the operating system allocated for the process/program, the process/program will not have the permission to read/write to that location, which will result in memory buffer location undefined.

Solution:

Python-based ATCRA services script is provided with the package. ATCRA services script requires the exact input as the main ATCRA; the only difference between ATCRA services and main ATCRA is that the services script will execute each individual PDF report, as its own process. In this way, the memory buffer will not be over flow and cause the memory address to be undefined.