

90th Transportation Research Board, Washington, D.C., USA, January 23-27, 2011.

# Investigating Collision Factors by Mining Microscopic Data of Vehicle Conflicts and Collisions

Nicolas Saunier, Ph.D. (corresponding author)

Assistant Professor, Department of Civil, Geological and Mining Engineering  
École Polytechnique de Montréal  
nicolas.saunier@polymtl.ca

Nadia Mourji

Graduate Student, Department of Mathematics and Industrial Engineering  
École Polytechnique de Montréal  
nadia.mourji@polymtl.ca

Bruno Agard, Ph.D.

Associate Professor, Department of Mathematics and Industrial Engineering  
École Polytechnique de Montréal  
bruno.agard@polymtl.ca

Word Count: (5167 Text) + (7 Figures) + (1 Table) = 7167 words

## **ABSTRACT**

Road collisions represent deplorable human and financial costs to society. Although some progress has been made, a renewed effort is necessary to tackle this growing worldwide issue. This paper advocates the development of proactive methods for road safety analysis that do not depend on the occurrence of collisions. In particular, collecting and analyzing microscopic data (road users' trajectories) about all traffic events with and without a collision is the only way to gain insight into collision factors and processes, i.e. the chains of events that lead to collisions. This paper reports on the first phase of a project relying on microscopic data extracted from video sensors and data mining techniques to identify patterns in the traffic event database. Decision trees, the k-means algorithm and a hierarchical agglomerative clustering method are used to analyze the relationship between interaction attributes and outcome (collision or not) and identify groups of interactions with similar attributes. This approach is demonstrated on a dataset collected in Kentucky of 295 traffic events, constituted of 213 conflicts and 82 collisions. The decision tree confirms the importance of the evasive action in the interaction outcome. Three clusters are found based on speed indicators extracted from the road users' trajectories: one cluster contains few collisions, with the lowest speeds among the three clusters. This result hints at the existence of conflicts that are dissimilar from most collisions and may therefore not be suitable for surrogate safety analysis.

## INTRODUCTION

The social cost of road collisions is among the largest negative side effects of road transportation. Including the costs of fatalities, disabilities, injuries and property damage, as well as medical care, lost productivity and traffic delays, the social cost of road collisions is estimated in (1) at \$63 billion for Canada in 2004. In many age intervals, road collisions are among the leading causes of death: in particular, it is the first leading cause of death for people between 15 and 29 years (2). The World Health Organization predicts that road collisions will jump from the ninth leading cause of death in 2004 to the fifth in 2030 (2). Compared to other health problems, the burden of road collisions is all the more acute because the victims are overwhelmingly young and healthy prior to their involvement in a collision.

Road safety improvements may be achieved within the three components of the road system through changes in

1. infrastructure design,
2. vehicle safety,
3. road user behaviour.

This work deals with methods for road safety analysis at a given location, which permit to identify contributing factors related to the three components and more particularly to the road. Traditional road safety analysis methods rely on collision databases that are filled with collision data manually collected after the occurrence of the collision, typically in the form of insurance and police reports. This data suffers from the following issues (3):

1. difficult attribution of collisions to a cause (reports are skewed towards the attribution of responsibility, not the search for the causes that led to a collision),
2. small data quantity,
3. limited quality of the data reconstituted after the event (with a bias towards more damaging collisions).

The following paradox ensues: safety analysts need to wait for accidents to happen in order to prevent them. There is a need for new proactive methods for road safety analysis, relying on more frequent traffic events without a collision.

Limited data is available on the context of collisions and the collision process, i.e. the chain of events that lead to a collision. The solution is to record information continuously about all traffic events: this can be achieved using video sensors and computer vision techniques to extract all road users' trajectories. Such data helps investigate the safety hierarchy (4), i.e. the framework that places all traffic events on a continuum with collisions at the top, undisturbed passages or "safe traffic events" at the bottom and traffic conflicts in between. The position of a traffic event in the safety hierarchy measures its proximity to a potential collision, or severity. Significant effort has been invested to develop techniques to collect and link to collisions the specific class of the most severe traffic conflicts. It is believed that the observation of all traffic events can provide a complementary safety diagnosis, more complete than can be done using collision data alone. It is in particular a way to gain more knowledge about the factors and processes that lead to collisions.

This work has implications for the development of proactive methods for road safety analysis. A better understanding of the characteristics of traffic events with and without a collision should help derive better relationships between them, in order to identify types of traffic events without a collision that can be used as surrogates for road safety analysis. This is critical as the work of Davis et al. (5) on a small set of traffic events suggests that the evasive actions undertaken by road users involved in conflicts may be of a different nature than the ones attempted in collisions.

This work pursues a line of research started at the University of British Columbia on automated road safety analysis using video sensors. A probabilistic framework for the computation of the probability of collision for all road users in interaction was proposed in (6) and refined in (7). This paper reports on the first phase of a research project that aims to better understand collision factors and processes using a large set of traffic events composed of conflicts and collisions. Contextual information and microscopic data, i.e. the trajectories of road users involved in the traffic events, are extracted automatically and mined for patterns using artificial intelligence techniques (data mining). By comparing traffic events with and without a collision, this work will help to identify surrogate measures of safety. To the authors' knowledge, this work is unique in the size of the analyzed dataset and the actual observation of safety related events. The next sections will cover: the related work, the descriptive analysis of the dataset, the results of the data mining techniques and finally the conclusion with a discussion of the next steps of the project.

## **RELATED WORK**

### **Road Safety Analysis**

Safety is defined as the number of collisions expected to occur at a given location per unit of time, where expected means "the average in the long run if it were possible to freeze all prevailing conditions that affect safety" (8). There has been a considerable amount of research to estimate safety models as a function of explanatory variables describing the transportation system: the road, the vehicle and the driver. These safety models, also called a crash prediction model (CPM) or safety performance function (SPF), typically take the form of an equation linking safety to a set of variables and rely on historical collision data. These models are at the core of the reference Highway Safety Manual (HSM) that should be available in 2010. As argued in the introduction, historical collision data obtained from insurance and police reports is ill-suited for the analysis of collision processes.

Two methods based solely on collisions stand out to help shed more light on collision factors and processes: in-depth accident analysis and naturalistic driving studies. In-depth accident analysis rely on detailed reconstitutions to investigate collision factors (9) and as such may provide some information on the chain of events that led to the collision. However, they share many shortcomings with methods based on historical collision data: they provide only limited amounts of data, at a higher cost, they rely on reconstitutions in which the collision processes may be only guessed at and they still require to wait for collisions to occur. Naturalistic driving studies rely on the continuous collection of data on a road user, his driving behaviour, the vehicle and the environment, over extended periods of time (10). Very large projects, e.g. in the Strategic Highway Research Program 2 Safety research area (11), are in the making and should provide unprecedented information. An advantage will be the observation of all traffic events, not only collisions. Nevertheless, naturalistic driving studies also have limitations: they typically provide detailed information only on one of the road users involved in a safety-related event; instrumenting

vehicles is costly and requires access to the vehicle, while fixed video cameras provide external non-intrusive monitoring of all traffic events and their context at a lower cost.

Traffic conflict studies are the most common proactive methods for road safety analysis (4, 12). Although mixed validation results, issues of cost and reliability have hindered their development, they have been integrated into traditional approaches, including the HSM, providing complementary information and alternative methods. The framework of the safety hierarchy was developed in the context of traffic conflict studies (13) and is the basis for more recent approaches that take into account all road users' interactions, not only the most severe traffic conflicts, for more complete and robust diagnoses (4, 6, 7). However, traffic conflict data traditionally collected by observers in the field is too limited and suffers from reliability issues that makes it unsuitable to understand collision processes: objective microscopic data is required for this purpose.

Limited road safety analysis is based on microscopic road user data as it was not easily available until recently: computer vision techniques now permit to extract automatically microscopic data from video data. Road users' trajectories are rarely collected with the primary goal of safety analysis (5, 14, 15, 16, 17, 18, 19). To the authors' knowledge, no work aiming to understand collision processes relies on automatically collected microscopic data. The present work is also unique in the size of the dataset and the fact that it contains traffic events with and without a collision.

### **Data Mining in Road Safety Analysis**

Machine learning models, like artificial neural networks (ANN) and support vector machines (SVM) (20), have been widely applied to estimate CPMs. However, the goal of this project is to understand collision factors, which requires to extract patterns from data and can be achieved through data mining techniques (21). These include classification, using for example decision trees that can be interpreted, as opposed to the "black box" nature of ANNs and SVMs, and clustering, i.e. finding groups through some similarity measure, using for example the k-means algorithm. Data mining has been used for the analysis of databases constituted only of collisions without any microscopic data.

Safety models for some collision attribute have been built to classify collisions and identify collision factors. Collision-prone locations and the membership of collision factors to road, driver of vehicle was investigated in (22, 23) using a fuzzy k-means algorithm, ANN and fuzzy k-nearest neighbours. Several attempts at modelling collision outcomes have been made. In (24), ANNs, decision trees and logistic regression are employed to identify collision severity-related factors to predict one of three possible outcomes: bodily injury (death or injury) and property damage. It was found that the presence of a protective device (i.e., seat-belt or helmet) is the most important factor in the crash severity variation. Further work in (25) indicate that a clustering-based classification algorithm works best for their data. ANNs and fuzzy adaptive resonance theory ANN are used in (26) to show that gender, vehicle speed, seat belt use, type of vehicle, point of impact, and area type (rural versus urban) affect the likelihood of injury severity levels. (27) developed classification and regression trees (CART) to analyze the risk factors that can influence the injury severity in traffic collisions: it was concluded that the most important variable associated with collision severity is the vehicle type, while pedestrians, motorcycle and bicycle riders are identified to have higher risks of being injured than other types of vehicle drivers in traffic collisions. Latent class clustering technique were employed in (28) to segment traffic collision data and identify homogeneous collision types. Injury models were then developed using multinomial logit model for each resulting

cluster, based on features such as collision type, crossroad type, built-up area, road type, road user age, dynamics of road user (moving or stationary), and vehicle type: it is concluded that cluster models reveal new variables affecting injury outcome and provide a more complete interpretation of the relationship causal variables and injury outcome. From this short review, it is apparent that, despite significant use of data mining techniques to analyze collision data, the lack of microscopic data describing traffic events with and without a collision limits the scope of the collision factors that can be identified and the analysis of the similarities of traffic events of different severities.

## **DESCRIPTIVE ANALYSIS OF THE DATASET**

### **Context**

This work relies on a unique dataset of video recordings of traffic conflicts and collisions collected at one signalized intersection Kentucky between August 16<sup>th</sup> 2001 to May 31<sup>st</sup> 2006 (29). All the analysis reported in this paper was carried out with the video recordings as the only source of information, except for the date and time of recording that can be derived from the filenames. There are two subsets of video recordings, one labelled as “miss” and the other as “incident”, corresponding respectively to traffic conflicts of mild to high severity and collisions. It is not clear from (29) how the severity was estimated to identify the subset of traffic conflicts. Each recording contains, or should contain, one clear safety-related traffic event, i.e. a traffic conflict or a collision. From the original set of respectively 238 traffic conflicts and 116 collisions, 213 and 82 are used in the analysis. The remainder of the recordings were not analyzed because of the video quality, of tracking issues for the video analysis tool or of the absence of a relevant traffic event. In the remaining video recordings, all safety-related traffic events involve at least two road users: all traffic events are therefore referred to as interactions.

The quality of the video data makes road user detection and tracking challenging. The video recordings have a resolution of 352 pixels in width by 240 pixels in height, varying levels of compression, colour aberrations..., affecting the image quality, and a frame rate of 15 frames per second. Many challenging conditions for automated video analysis are covered, with various times of recording (day and night) and weather conditions: sunny days cause strong shadows, there are many cases of snow, fog and rain (sometimes at night, in which case the reflection of vehicle headlights causes particular glare). Although these issues made some recordings impossible to analyze, road users’ detection and tracking was possible in most recordings, using a video-based system developed previously (30) (see some frames in FIGURE 4). The analysis relies on the trajectories used in (7).

### **The Database**

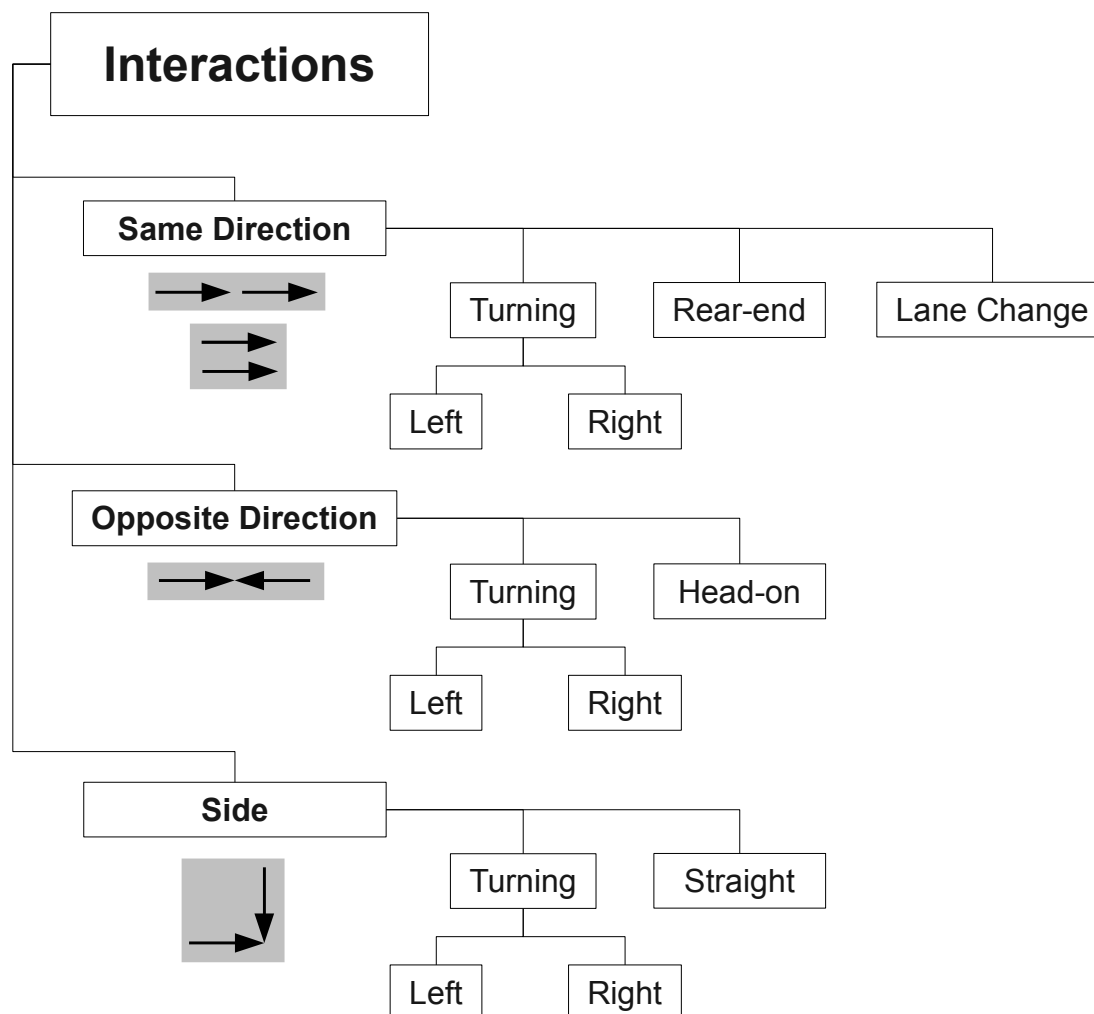
All 295 interactions are described by a list of attributes, listed in TABLE 1. Most categorical attributes are extracted manually by watching the video recordings. The simple interaction categories used in (7) are expanded in this work (see FIGURE 1), covering most of the categories proposed in (8). The road users’ origin and type are also stored in the database: the number of road users involved in each interaction is counted by type. In addition to bikes and pedestrians, the types of motorized vehicles are taken from the FHWA vehicle classes. There are currently no pedestrians, bikes or buses in the database. In some cases, more than two road users could be considered to be involved in an interaction, but for simplicity in this first phase, information only on the two closest was included in the database. The number of evasive actions attempted by the road users are also counted by type.

**TABLE 1** Attributes of the interactions (A  $\times$  in the “Auto” column indicates if the attributes were automatically extracted from the data,  $\Delta v$  is the times series of the norm of the difference of the velocities, and  $s$  designates the road users’ speed time series)

<b>Categorical attributes</b>	<b>Auto</b>	<b>Values</b>
<i>Type of day</i>	$\times$	weekday, week end
<i>Lighting condition</i>		daytime, twilight, nighttime
<i>Weather condition</i>		normal, rain, snow
<i>Interaction category</i> (FIGURE 1)		same direction (turning left and right, rear-end, lane change), opposite direction (turning left and right, head-on), side (turning left and right, straight)
<i>Interaction outcome</i>		conflict, collision
<b>Numerical attributes</b>	<b>Auto</b>	<b>Units</b>
<i>Road user type</i> passenger car van, 4x4, SUV bus truck (all sizes) motorcycle bike pedestrian		number of road users number of road users number of road users number of road users number of road users number of road users number of road users
<i>Road user origin</i> I65 Jefferson Street exit I65 Brook Street exit Brook Street Jefferson Street		number of road users number of road users number of road users number of road users
<i>Type of evasive action</i> No evasive action Braking Swerving Acceleration		number of evasive actions number of evasive actions number of evasive actions number of evasive actions
<i>3 attributes from <math>\Delta v</math></i>	$\times$	km/h
<i>6 values from <math>s</math></i>	$\times$	km/h

To extract automatically information from the trajectories of the road users involved in each interaction, they are manually identified among all road users’ trajectories in each video sequence. For this phase of the project, only speed data was extracted: from the times series of the road users velocities (speed vectors), the following attributes are computed:

- the minimum, maximum and mean values of the speeds of each road user, denoted respectively  $s_{min}$ ,  $s_{max}$  and  $\bar{s}$ . In order to have a unique description of the interaction, i.e. symmetric with respect to the involved road users, the attributes are ordered by increasing value (i.e.  $s_{min1} < s_{min2}$ ,  $s_{max1} < s_{max2}$  and  $\bar{s}_1 < \bar{s}_2$ ). 6 such attributes are therefore used to describe the road users’ speeds during their interaction.
- the minimum, maximum and mean values of the norm of the difference of the road users’



**FIGURE 1 Hierarchy of interaction categories.**

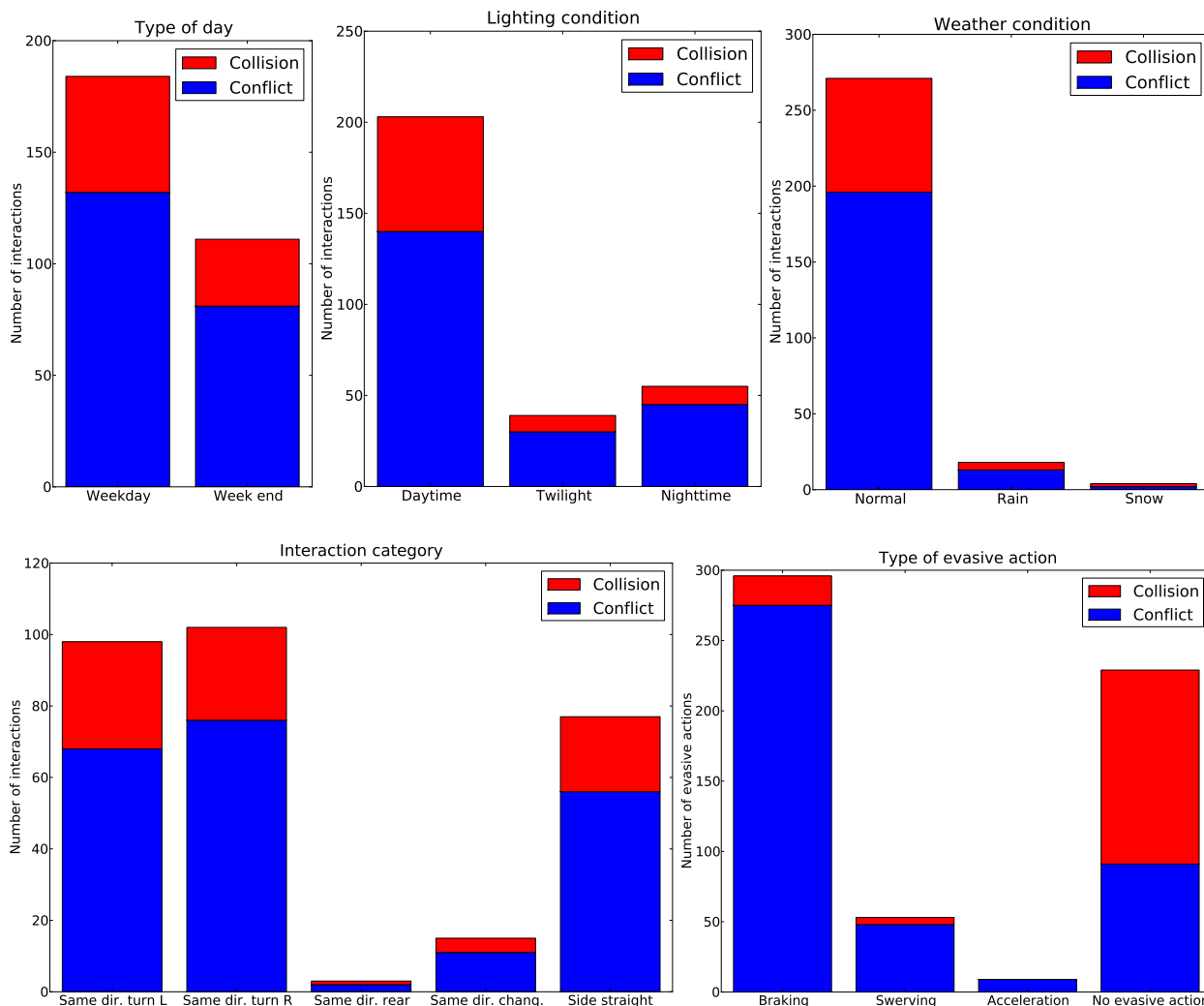
velocities for the whole time interval during which they are tracked simultaneously, denoted respectively  $\Delta v_{min}$ ,  $\Delta v_{max}$  and  $\overline{\Delta v}$ .

### Descriptive Analysis

As can be seen in FIGURE 2, there does not seem to be any relationship between the interaction outcome and the type of day, the lighting conditions and the weather conditions. Overall, weather conditions were mostly “normal”. There are actually slightly fewer collisions during twilight and nighttime than during daytime. There is also no particular pattern in the interaction category. As can be expected, there is a strong relationship between the type of evasive action and the interaction outcome: in most collisions (62 out of 82), no evasive action was attempted. There is also a sizable amount of conflicts where at least one of the involved road users did not attempt an evasive action.

Speed attributes are displayed in FIGURE 3. The speed differences are quite similar for conflicts and collisions. However, speeds are systematically higher for collisions than for conflicts, although the difference is within the standard deviation of each category.



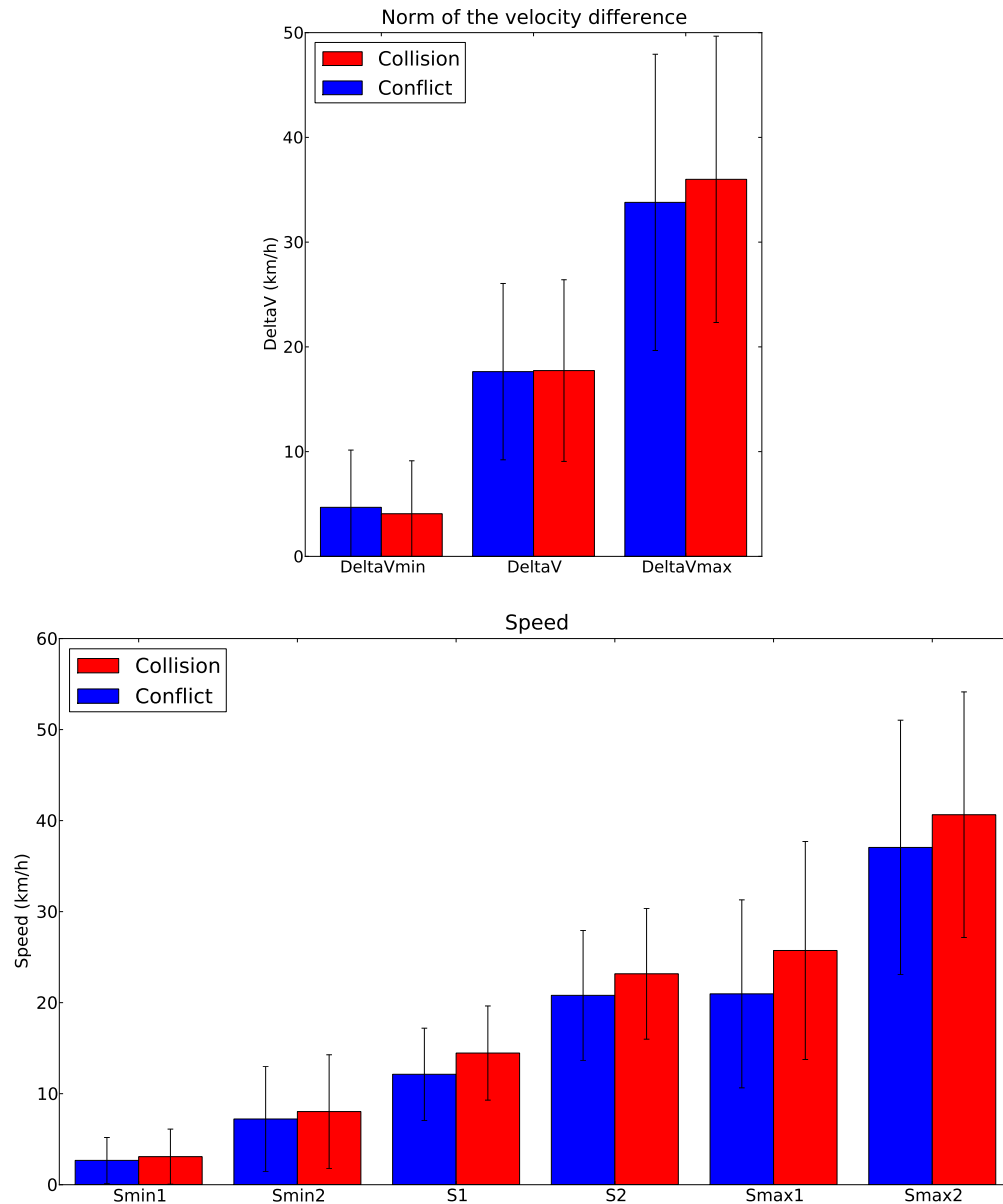


**FIGURE 2** Some distributions of the database attributes per interaction outcome.

A rough localization of each interaction was also obtained by averaging the positions of the involved road users during the period of time they were simultaneously tracked. No pattern was visible when plotting the various interaction attributes, in particular the interaction outcome, except for the interaction category (see FIGURE 4). This makes sense as the side straight interactions occur at the upper left corner of the intersection, while same direction interaction happen between vehicles coming from similar origins on either street.

### EXPLORATORY ANALYSIS USING DATA MINING TECHNIQUES

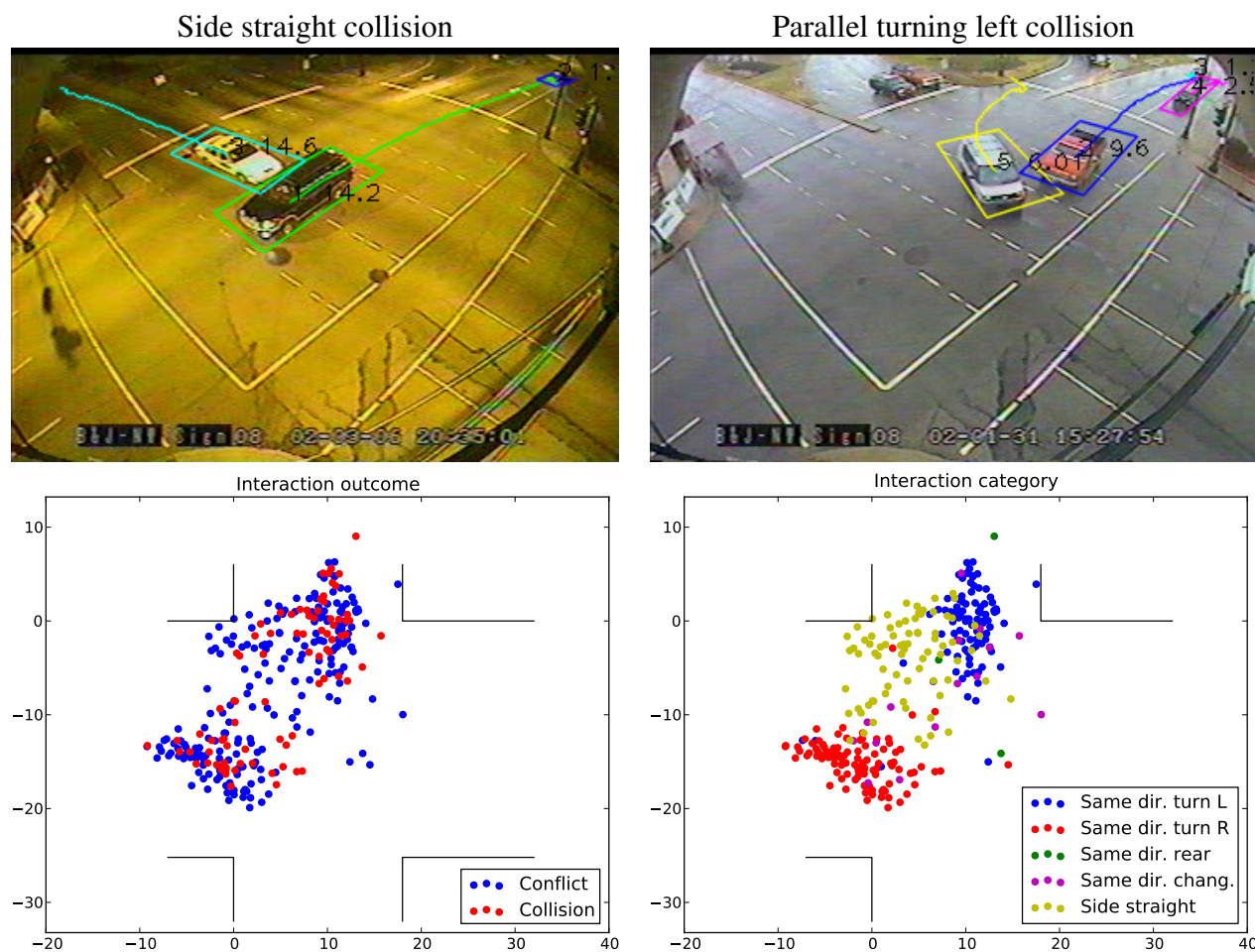
The interaction database was mined for patterns using two well-known data mining techniques: decision trees and the k-means algorithm (21). Association rules were tried, but did not yield any strong result. The free and open source software TANAGRA is used for this analysis (31).



**FIGURE 3** Some distributions of the speed attributes of the database per interaction outcome (the error bars represent the standard deviation).

### Classification

The C4.5 decision tree algorithm is used to predict the interaction outcome from the attributes. The goal is not to classify interactions, but to identify rules in the conditions generated by the algorithm at each node to predict the interaction outcome. The following numerical attributes are transformed and grouped as categorical attributes so that they are treated as one attribute by the decision tree: the road users types, the road users origins and the types of evasive actions. The new attribute can be any pair of two of the original attributes, e.g. “passenger car/truck” for the road user types or “braking/swerving” for the types of evasive actions. There is no need to normalize



**FIGURE 4** Sample interactions of different categories (top) and spatial plots of the interaction outcomes (bottom left) and categories (bottom right). The upper left intersection corner in the maps (bottom) corresponds to the farthest corner in the video images (top).

numerical attributes for decision trees.

A decision tree is thus learnt using all the interaction attributes, with the parameters minimum size of leaves set to 6 and the confidence level to 0.25: the decision tree has 13 nodes and 10 leaves, and an error rate of 0.0746 (which is computed on the training dataset and therefore over-estimates the classification performance). The resulting rules are presented in FIGURE 5, with the characterization of the corresponding leaf. The first split done by the tree deals with the evasive actions, which is expected (see FIGURE 2): in particular, the absence of an evasive action leads in 91.18% of cases to a collision. The presence of at least an evasive action is associated with conflicts, except for “swerving/no evasive action” and “braking/no evasive action”. In this latter case, it is possible to refine the rules based on the speed attributes  $\overline{\Delta v}$  and  $\bar{s}_1$ . Conflicts are associated with larger mean velocity differences, or, if the average velocity difference is low, lower mean individual speeds: the first case is probably related to successful braking to avoid collision, while the second is more difficult to interpret. Collisions occur in the remainder of the cases, i.e. for low mean velocity differences and higher individual mean speed, which is logical, though pertains to

- Evasive actions in [braking/braking] then Interaction outcome = conflict (94.62 % of 93 examples)
- Evasive actions in [braking/no evasive action]
  - $\overline{\Delta v} < 12.6183$ 
    - \*  $\bar{s}_1 < 13.4022$  then Interaction outcome = conflict (83.33 % of 12 examples)
    - \*  $\bar{s}_1 \geq 13.4022$  then Interaction outcome = collision (83.33 % of 6 examples)
  - $\overline{\Delta v} \geq 12.6183$  then Interaction outcome = conflict (95.31 % of 64 examples)
- Evasive actions in [no evasive action/no evasive action] then Interaction outcome = collision (91.18 % of 68 examples)
- Evasive actions in [braking/swerving] then Interaction outcome = conflict (96.55 % of 29 examples)
- Evasive actions in [no evasive action/swerving] then Interaction outcome = conflict (55.56 % of 9 examples)
- Evasive actions in [swerving/swerving] then Interaction outcome = conflict (100.00 % of 5 examples)
- Evasive actions in [braking/acceleration] then Interaction outcome = conflict (100.00 % of 7 examples)
- Evasive actions in [no evasive action/acceleration] then Interaction outcome = conflict (100.00 % of 2 examples)

**FIGURE 5 Rules generated by the decision tree.**

only 5 interactions out of 6.

This analysis has provided interpretable knowledge about the interactions and their attributes and confirms the obvious link between collision avoidance and the presence of at least an evasive action. In particular, mean velocity differences are higher when collision is avoided thanks to braking by one of the road users.

### **Clustering**

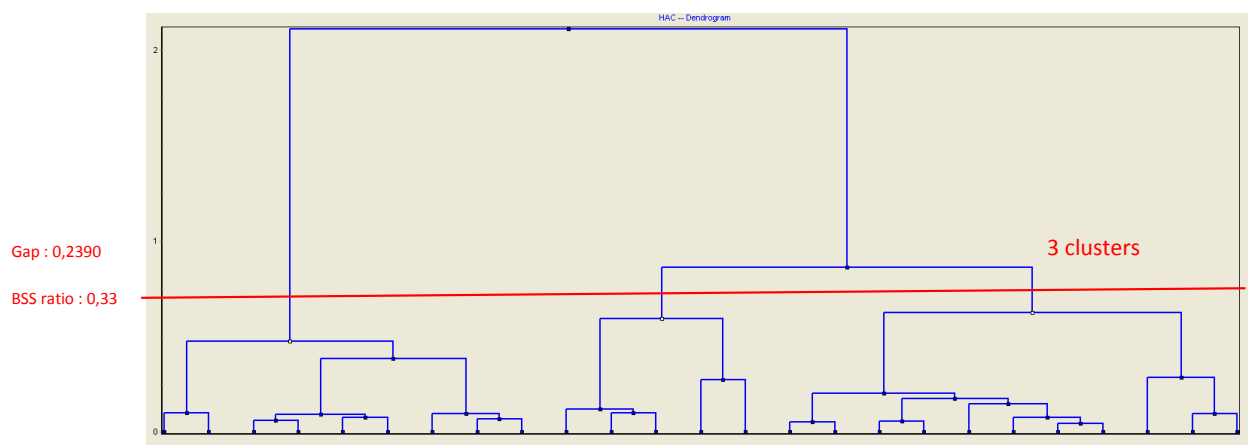
Given the descriptive analysis of the database done in the previous section and the goal of this project to study collision processes, the choice was made to use only the nine speed attributes for the clustering. In particular, it is hoped that the resulting clusters can help identify relationships between conflicts and collisions that can be used for surrogate safety analysis, as well as the lack of such relationships. Three types of clusters can be produced by the method, depending on the proportion of conflicts and collisions in the cluster, with the following potential implications:

- a mixed cluster of similar conflicts and collisions: this could indicate that the conflicts in the cluster can be used as surrogates for the collisions in the same cluster,

- a pure cluster with no or very few conflicts in the cluster: this could indicate that it is not possible to use conflicts as surrogates to the collisions in the cluster,
- a pure cluster with no or very few collisions in the cluster: this could indicate that the conflicts in the cluster cannot be used as surrogates to any type of collision.

Before doing any analysis, the required preliminary step is to normalize all the attributes: the objective is to standardize the scale of effects of each variable on the results. The Euclidean distance is chosen as the distance used by the k-means algorithm to compare the interaction instances.

The partitioning algorithm k-means is a very effective method used to identify homogeneous groups assuming a number of classes known at the beginning. The number of clusters depends on the depth of analysis desired and its determination can be a challenge. Following the approach used in (32), the data is first clustered in a large number of groups, larger than what can be reasonably expected for our analysis (in our case 25). The groups are then merged using a hierarchical agglomerative clustering method. At each iteration, the ratio of between-cluster sum-of-squares (BSS ratio) and the Gap are calculated and used to determine the number of clusters by the hierarchical algorithm. The compactness of the data is measured by the gap value while the dissimilarity between them is considered by the BSS ratio. A good clustering yields clusters where they have high BSS ratio and gap value. The dendrogram shows that a division into three clusters is appropriate (see FIGURE 6).



**FIGURE 6 Dendrogram of the 25 clusters to obtain the number of clusters in the data.**

After identifying an appropriate number of clusters with the previous method, the k-means algorithm is applied again for three groups. Some distributions of the clusters are presented in FIGURE 7. Cluster 1 and 3 are characterized by higher proportions of collisions (respectively 40.8 % and 44.4 %), while cluster 2 contains few collisions (7.8 %). The distribution of evasive actions per cluster is not shown, but is consistent with the previous characterization and the proportions of conflict and collisions in the clusters. The distribution of interaction categories in cluster

2 is homogeneous to the whole database, while it is skewed in cluster 1 (over-representation of side straight and same direction turning right) and cluster 3 (over-representation of same direction turning left and right and same direction changing lanes). No relationship to the other categorical attributes could be discovered. Since the clustering is based on speed attributes, the clusters should show differences for these attributes. As can be seen in FIGURE 7, the clusters comprehend different mean speed and velocity difference values: the first cluster has the highest speeds for all attributes, followed by cluster 3 and cluster 2 on all attributes but one. Similarly to FIGURE 4, the spatial distribution of the clusters was investigated but showed no discernible pattern. The final characterization of the clusters, i.e. of the over-representations of some attributes with respect to the whole database, can thus be made:

**Cluster 1** collisions, highest speeds, categories side straight and same direction turning right

**Cluster 2** almost pure conflicts, lowest speeds

**Cluster 3** collisions, medium speeds, categories same direction turning left and right and same direction changing lanes

Based on a clustering of the interactions only on speed attributes, the three clusters show some contrasted characteristics. The clusters 1 and 3 are mixed, with over-representation of collisions with respect to the whole database, and represent two classes of interactions differing in speeds and categories: the conflicts in these clusters are good candidates for surrogates for the collisions in their respective clusters. On the other hand, cluster 2 is overwhelmingly constituted of conflicts, and characterized by lower speeds: it could characterize a group of conflicts of lower severity, that cannot be used as surrogates, or only for some specific collisions.

## CONCLUSION

This paper has presented the first results of a larger project aiming to better understand collision factors and processes using microscopic data. A large dataset of 295 interactions, constituted of conflicts and collisions, was characterized by mining their attributes, using in particular decision trees and the k-means partitioning algorithm. Obvious relationships such as the link of evasive actions and their absence to the interaction outcome were confirmed. The clustering analysis yielded evidence that not all conflicts should be used as surrogates for all collisions and showed how groups of similar conflicts and collisions can be identified.

Further investigation is necessary to confirm these findings. It should be recalled that the conditions of the data collection are not exactly known, in particular their sampling conditions: is the set of interactions available in the dataset representative of all interactions of similar severity that occurred during the time period of the data collection? It should also be noted that exposure data is not available and that no conclusion should be therefore drawn on the risk of collision in the conditions under study.

In any case, this work paves the way for larger efforts to come and increase knowledge of collision processes. The next phases of this work will make use of the whole road users' trajectories, develop temporal indicators to characterize the interactions and use better similarities, which can in particular accommodate multi-dimensional vectors of varying length. Future projects will include the collection of large scale datasets of all road users' interactions in known conditions that will allow to draw stronger conclusions.

## ACKNOWLEDGEMENTS

The authors wish to acknowledge the financial support of the Fondation Polytechnique and the Research and Innovation Directorate of École Polytechnique of Montréal, and to thank Zu Kim of California PATH and Ann Stansel of the Kentucky Transportation Cabinet for providing the video dataset.

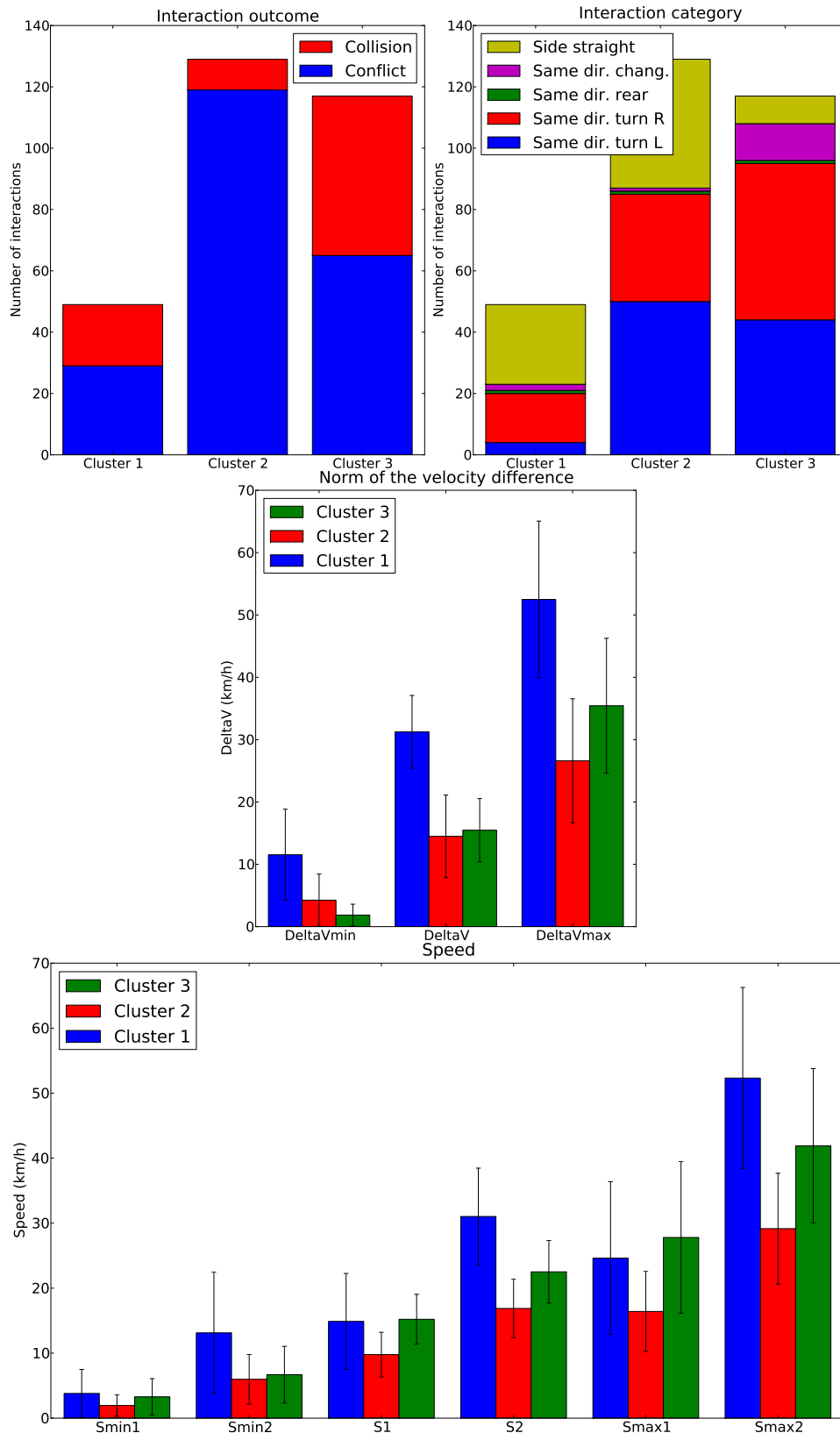
## REFERENCES

- [1] Vodden, K., D. Smith, F. Eaton, and D. Mayhew, *Analysis and Estimation of the Social Cost of Motor Vehicle Collisions in Ontario*. Final Report TP 14800F, Transport Canada, 2007.
- [2] World Health Organization, *Global status report on road safety: time for action*. Geneva, 2009.
- [3] Ismail, K., T. Sayed, and N. Saunier, Automated Analysis Of Pedestrian-vehicle Conflicts: A Context For Before-and-after Studies. In *Transportation Research Board Annual Meeting Compendium of Papers*, Washington, D.C., 2010, 10-3739. Accepted for publication in *Transportation Research Record: Journal of the Transportation Research Board*.
- [4] Svensson, A. and C. Hydén, Estimating the severity of safety related behaviour. *Accident Analysis & Prevention*, Vol. 38, No. 2, 2006, pp. 379–385.
- [5] Davis, G. A., J. Hourdos, and H. Xiong, Outline of Causal Theory of Traffic Conflicts and Collisions. In *Transportation Research Board Annual Meeting Compendium of Papers*, 2008, 08-2431.
- [6] Saunier, N. and T. Sayed, A Probabilistic Framework for Automated Analysis of Exposure to Road Collisions. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2083, 2008, pp. 96–104.
- [7] Saunier, N., T. Sayed, and K. Ismail, Large Scale Automated Analysis of Vehicle Interactions and Collisions. In *Transportation Research Board Annual Meeting Compendium of Papers*, Washington, D.C., 2010, 10-4059. Accepted for publication in *Transportation Research Record: Journal of the Transportation Research Board*.
- [8] Hauer, E., J. Ng, and J. Lovell, Estimation of Safety at Signalized Intersections. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1185, 1988, pp. 48–61.
- [9] Fleury, D. and T. Brenac, Accident prototypical scenarios, a tool for road safety research and diagnostic studies. *Accident Analysis & Prevention*, Vol. 2, No. 33, 2001, pp. 267–276.
- [10] Neale, V., T. Dingus, S. Klauer, J. Sudweeks, and M. Goodman, An overview of the 100 car naturalistic study and findings. In *Proceedings of the 19th ESV Conference*, Washington, D.C., 2005.
- [11] Committee for the Strategic Highway Research Program 2: Implementation, *Implementing the Results of the Second Strategic Highway Research Program: Saving Lives, Reducing Congestion, Improving Quality of Life*. Special Report 296, Transportation Research Board, 2009.

- [12] Sayed, T. and S. Zein, Traffic conflict standards for intersections. *Transportation Planning and Technology*, Vol. 22, 1999, pp. 309–323.
- [13] Hydén, C., *The development of a method for traffic safety evaluation: The Swedish Traffic Conflicts Technique*. Ph.D. thesis, Lund University of Technology, Lund, Sweden, 1987, bulletin 70.
- [14] Kamijo, S., Y. Matsushita, K. Ikeuchi, and M. Sakauchi, Traffic monitoring and accident detection at intersections. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 2, No. 2, 2000, pp. 108–118.
- [15] Hu, W., X. Xiao, D. Xie, T. Tan, and S. Maybank, Traffic Accident Prediction using 3D Model Based Vehicle Tracking. *IEEE Transactions on Vehicular Technology*, Vol. 53, No. 3, 2004, pp. 677–694.
- [16] Messelodi, S. and C. M. Modena, *A Computer Vision System for Traffic Accident Risk Measurement: A Case Study*. ITC, 2005.
- [17] Atev, S., H. Arumugam, O. Masoud, R. Janardan, and N. P. Papanikolopoulos, A vision-based approach to collision prediction at traffic intersections. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 6, No. 4, 2005, pp. 416–423.
- [18] Lareshyn, A. and H. Ardö, Automated Video Analysis as a Tool for Analysing Road Safety Behaviour. In *ITS World Congress*, London, 2006.
- [19] Kim, Z., Real time object tracking based on dynamic feature grouping with background subtraction. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.
- [20] Xie, Y., D. Lord, and Y. Zhang, Predicting motor vehicle collisions using Bayesian neural network models: An empirical analysis. *Accident Analysis & Prevention*, Vol. 39, No. 5, 2007, pp. 922–933.
- [21] Duda, R. O. and P. E. Hart, *Pattern Classification*. Wiley-Interscience, 2000.
- [22] Sayed, T., W. Abdelwahab, and F. Navin, Identifying Accident-Prone Locations Using Fuzzy Pattern Recognition. *Journal of Transportation Engineering*, Vol. 121, No. 4, 1995, pp. 352–358.
- [23] Sayed, T. and W. Abdelwahab, Comparison of Fuzzy and Neural Classifiers for Road Accidents Analysis. *Journal of Computing in Civil Engineering*, Vol. 12, No. 1, 1998, pp. 42–47.
- [24] Sohn, S. Y. and H. Shin, Pattern recognition for road traffic accident severity in Korea. *Ergonomics*, Vol. 44, No. 1, 2001, pp. 107–117.
- [25] Sohn, S. Y. and S. H. Lee, Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea. *Safety Science*, Vol. 41, No. 1, 2003, pp. 1–14.



- [26] Abdel-Aty, M. A. and H. T. Abdelwahab, Predicting Injury Severity Levels in Traffic Crashes: A Modeling Comparison. *Journal of Transportation Engineering*, Vol. 130, No. 2, 2004, pp. 204–210.
- [27] Chang, L.-Y. and H.-W. Wang, Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis & Prevention*, Vol. 38, No. 5, 2006, pp. 1019–1027.
- [28] Depaire, B., G. Wets, and K. Vanhoof, Traffic accident segmentation by means of latent class clustering. *Accident Analysis & Prevention*, Vol. 40, No. 4, 2008, pp. 1257–1266.
- [29] Green, E. R., K. R. Agent, and J. G. Pigman, *Evaluation of Auto Incident Recording System*. Kentucky Transportation Cabinet, 2005.
- [30] Saunier, N. and T. Sayed, A feature-based tracking algorithm for vehicles in intersections. In *Third Canadian Conference on Computer and Robot Vision*, IEEE, Québec, 2006.
- [31] Rakotomalala, R., TANAGRA : un logiciel gratuit pour l'enseignement et la recherche. In *Actes de EGC'2005*, 2005, Vol. 2, pp. 697–702.
- [32] Le, T., B. Agard, and S. Deveault, Decision support based on socio-demographic segmentation and distribution channel analysis in the US furniture market. In *International Conference on Industrial Engineering and Systems Management - IESM' 2009*, Montréal, Canada, 2009.



**FIGURE 7** Distributions of the interaction outcome, interactions category, and speed attributes for the three clusters (the error bars represent the standard deviation).