Clustering Surrogate Safety Indicators to Understand Collision Processes

Nicolas Saunier (corresponding author)
Assistant Professor, Department of Civil, Geological and Mining Engineering
Polytechnique Montréal
nicolas.saunier@polymtl.ca

Mohamed Gomaa Mohamed
PhD Candidate, Department of Civil, Geological and Mining Engineering
Polytechnique Montréal
mohamed.gomaa@polymtl.ca

5333 words + 8 figures + 1 tables
August 1, 2013

ABSTRACT

As time series are collected through more and more pervasive devices carried by users and vehicles, new tools are necessary to understand and mine the large amounts of transportation data being thus generated. This work proposes a new similarity measure for time series that is applied to surrogate measures of safety and other indicators characterizing road user interactions. The new similarity measure based on the aligned longest common sub-sequence is paired with a custom clustering algorithm that does not require to set the number of expected clusters and remains interpretable through the use of prototype indicator profiles as cluster representatives. The method is applied to five indicators, including time to collision and probability of collision, for a large real world dataset of traffic videos of collisions and conflicts. The results confirm the general assumption of surrogate methods for safety analysis that some interactions without a collision have very similar processes to collisions. It also highlights the danger of using a significant proportion of candidate interactions without a collision that seem to share little similarities with collisions.

INTRODUCTION

As Moore's law makes sensors and computers ubiquitous in vehicles, road environments and on users, more and more data is collected continuously on all vehicles and road users. Examples are location data from vehicle and personal GPS sensors, road user trajectories extracted from traffic cameras, vehicle kinematic data and engine operational data from on-board diagnostics (OBD) devices. Storing this data results in large datasets of temporal measurements characterizing different elements of the road system. This data has the potential to be useful for several transportation applications, e.g. activity patterns, vehicle-based and site-based safety diagnosis, calibration and validation of macroscopic and microscopic models, behaviour observations at various space and time scales. However the promise of this "big data" can only be fulfilled if new methods are developed to deal with it and mine the large datasets that can be accumulated. Though aggregating spatio-temporal data over time and space wastes the potential of data collected at much finer resolutions, analysis often rely on reduced data for lack of expertise and tools, and for practical reasons.

Of particular interest is the development of methods for the surrogate analysis of safety. Traditional collision-based diagnosis methods have several shortcomings that have been repeatedly covered in previous work, e.g. in (*1, 2*). There is therefore a search for proactive methods that do not require to wait for accidents to occur. These surrogate methods rely on the observation of all interactions and the measure of their "severity" or proximity to a potential collision through continuous safety indicators such as the time to collision (TTC). These observations are more and more commonly obtained automatically through vehicle-based sensors such as data loggers (*3*) or dedicated devices installed for example for naturalistic driving studies (*4*) and site-based sensors such as video cameras with video analysis software (*2, 5, 6*). However, most analyses of this data and, to the authors' knowledge, all analyses of surrogate safety indicators rely on the aggregation of the temporal indicators into a single value. The most commonly used in traffic conflict analyses is the minimum TTC, or a severity level based on the TTC at a specific instant coupled with the road user speed at the same instant in the case of the Swedish traffic conflict technique (*7*). This is a terrible loss of information that could partially explain the mixed results to validate and transfer surrogate measures of safety. As stated in (*8*), "the problem with values taken at a certain time is that they do not incorporate any information before or after the chosen moment, creating a risk that even very different encounters might be classified in the same category".

This paper is a follow up on (*9*) that relied on contextual information and aggregated measures of road users' individual speeds and speed differential to cluster interactions with and without a collision. The purpose is to better understand collision processes and the similarities between all interactions. This will help determine whether all interactions without a collision can be used as surrogates for collisions. There is preliminary evidence that this is not the case and that some categories of interactions or interactions of different severity levels may not be associated with safety, i.e. that more interactions would translate into more collisions over the long run, and even be indicators of a good level of safety through the promotion of driver awareness and learning through interactions with other road users (*10*).

This paper presents ongoing work on the development of a method to compare and cluster time series or profiles of interaction indicators, including surrogate measures of safety. A new similarity measure based on the longest common sub-sequence (*11*) is proposed to better measure indicator profile similarity by taking into account the rate of change. A custom clustering algorithm is developed that does not require to set the number of expected clusters and remains interpretable

through the use of prototype indicator profiles as cluster representatives. The method is demonstrated on a large video dataset of interactions with and without a collision. The final contribution of this paper is the development of an open source library that implements the proposed methods and the release of the exact code used to produce the presented analyses in order to encourage reproducibility and wider adoption of the methods. The background is presented in the next section. It is followed by a description of the method, which is then demonstrated on a real world dataset. Finally the paper is concluded and future work is discussed.

BACKGROUND

Surrogate Safety Analysis

There is a growing body of literature on surrogate methods for safety analysis and readers are referred to these PhD theses (*12, 13, 14*) and the TRB white paper (*15*) for an introduction to the field and a coverage of the early techniques known as traffic conflict techniques (TCT). The defining characteristic of relevant traffic events for safety is the collision course, which is the situation in which two road users would collide if their movements remain unchanged (taken from the definition of a traffic conflict as "an observable situation in which two or more road users approach each other in time and space to such an extent that there is a risk of collision if their movements remain unchanged" (*16*)). Identifying a collision course at a given instant therefore requires to predict road users' future positions from their current and past positions. The default motion prediction method is to assume that the road users will move at constant velocity. The choice is rarely justified, does not yield robust measurements and does not take the context (the road, e.g. in a curve, and traffic) into account which results in unrealistic motion prediction (e.g. going off the road or into a wall). New prediction methods have been proposed in (*17, 18*) with open source implementations (*6*).

For surrogate safety analysis to be objective, a number of quantitative safety indicators have been proposed in the literature to measure the proximity to a potential collision, or probability of collision, and the severity of the potential collision. TTC is the best known of these indicators. It is defined for a given motion prediction method as the time required for two road users to collide following the predicted trajectories. If several predicted trajectories are available, with corresponding probabilities, the expected TTC can be computed (*2*). Many other safety indicators, including post-encroachment time (PET), deceleration to safety time, etc., have been presented over the years (see (*12, 13, 14*) and their references for more details).

Interpreting Interactions and Safety Indicators

In most TCTs, a specific value of a continuous safety indicator is used and compared to a threshold to distinguish, usually for diagnosis purpose, the most severe conflicts from "safer" interactions, defined as a situation in which two road users are within some distance. For example, Hydén (*19*) used the TTC just before one of the road users attempts an evasive action called the time to accident with a threshold of 1.5 s to define severe conflicts. The Federal Highway Administration (FHWA) designed the piece of software Surrogate Safety Assessment Model (SSAM) to perform the analysis of trajectory data extracted from microscopic simulation software (*20*). SSAM uses a predefined threshold for different safety indicators to identify the most severe conflicts among all road user interactions (e.g. the default threshold on minimum TTC is 1.5 s). The most severe value of safety indicators is typically used to summarize them, for example minimum values for spatio-temporal indicators (e.g. distance, TTC and predicted PET) or maximum values for probability

of collision or deceleration to safety time. However, as argued in the introduction and in (*8*), narrowing down the whole interaction to a single value leads to losing a lot of information. Even the work of Minderhoud and Bovy (*21*) highlighted in (*8*) still condenses the whole indicator profile into a single measure through integration.

There are few examples of the use or interpretation of continuous traffic event indicators over certain time intervals. Some studies on driver behaviour have relied on speed profiles. Parkhurst (*22*) examined the shape of speed profiles to understand the driver behaviour at urban and rural non-signalized intersections. Laureshyn et al. (*23*) classified the speed profiles, extracted using automated video analysis, of vehicles making left turns at a signalized intersection and interacting with oncoming traffic and crossing pedestrians. Among the three types of pattern recognition techniques tested, cluster analysis (k-means), supervised learning (k-nearest neighbours), and dimension reduction, k nearest neighbours was found to perform well with respect to the human observer annotations.

The goal of using whole safety indicators time series is to better understand collision processes and how interactions with and without a collision compare. Indeed, the work of Davis et al. (*24*) 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. The work of Svensson and Hydén (*10*) provides some evidence that interactions with fairly high severities could be associated with improved safety because they are frequent and severe enough to create and maintain awareness among road users.

Time-Series Clustering

The objective of clustering is to classify the data into groups (clusters) with similar characteristics. Because the groups are not known, clustering is also called unsupervised classification. Many algorithms have been proposed in the machine learning literature (*25*), e.g. hierarchical, based on density, centroids, statistical distributions, etc. A time series is a data type that represents a sequence of observation vectors $X(t) = [x_1(t), ..., x_n(t)]$ as a function of time t, usually at discrete instants. Time series can be univariate (one variable per observation, $n = 1$) or multivariate (many variables per observation, $n \geq 2$). The readers are referred to (*26, 27*) for surveys of clustering methods for time series data. Among the different algorithms developed in various domains, most attempt to reduce the dimensionality of the data to enhance the clustering performance. For example, Vlachos et al. (*28*) used k-means clustering incrementally at different levels (resolution) based on discrete wavelet transformation (DWT) decomposition.

The key component of many clustering methods is the measure of the similarity or distance between pairs of elements in the series. The Euclidean distance is popular, but it requires that both time series have the same length and it is sensitive to distortions (e.g. shifting along the time axis) and noise. The development of elastic distance and similarity measures, such as dynamic time warping (DTW) and longest common sub-sequence similarity (LCSS), overcome the previous drawbacks. Both DTW and LCSS are implemented using dynamic programming. DTW attempts to find the best alignment between two time series by minimizing the distance between them. Conversely, LCSS finds the length of the longest matching sub-sequence by comparing every point of the two time series using a given matching method. Morris and Trivedi (*29*) evaluated different similarity measures (HU, PCA (Principle Component Analysis), DTW, LCSS, PF (Piciarelli and Foresti (*30*)), Modified Hausdorff) and clustering methods for trajectories as a first step to understand road user behaviour. After tests on six different datasets, the authors concluded that

LCSS was consistently the top performer.

A relevant example for transportation of multivariate time series clustering is trajectory clustering which is done in surrogate safety analysis (*17*) and robotics applications. The objective is to cluster a dataset of observed trajectories into the main motion patterns. Bennewitz et al. (*31*) learnt the motion patterns of people in a scene using the Expectation Maximization (EM) algorithm, which enabled a robot to update its behaviour accordingly. Hu et al. (*32*) modelled road user activities with a fuzzy self-organizing neural network. One of the main applications is future motion prediction. Recently Morris and Trivedi (*33*) proposed a 3-stages hierarchical learning framework to analyze object activities and to predict future activities, as well as to detect abnormal events. The authors used LCSS as a similarity measure and spectral clustering algorithm (*34*) for the trajectory clustering.

PROPOSED APPROACH

Real time series from transportation will have varied lengths. This is the case for the analysis of safety indicators investigated in this paper. The choice is made to avoid pre-processing the data that would introduce distortions and may lead to loss of information for example through re-sampling. Methods that can deal with the data as it is, without pre-processing, are therefore preferred. Among the various such methods, the LCSS is favoured as it is flexible and can be adapted to specific purposes.

The Aligned Longest Common Sub-sequence

Let $X = [X(t_1), ...X(t_n)]$ and $Y = [Y(t_1), ...Y(t_m)]$ be two time series of respective length n and m of safety indicators characterizing two interactions (the series may be multivariate, e.g. if concatenating several indicator measurements at each instant). Let $Head(X)$ be the series $[X(t_1), ...X(t_{n-1})]$. Given a real number $\delta > 0$ and a matching function *match* for the elements of the series (e.g. for univariate series and a given real number $\varepsilon > 0$, $d_\varepsilon(a,b)$ is $true$ if $|a-b| \leq \varepsilon$, $false$ otherwise), the length $LCS_{\delta,match}(X,Y)$ of the longest common sub-sequence is computed as

- 0 if $m = 0$ or $n = 0$,
- $1 + LCS_{\delta,match}(Head(X), Head(Y))$ if $match(X(t_n), Y(t_m))$ is $true$ and $|n-m| \leq \delta$,
- $max(LCS_{\delta,match}(Head(X), Y), LCS_{\delta,match}(X, Head(Y)))$ otherwise.

This is typically computed in a matrix S using dynamic programming where $S_{i,j}$ is the *LCS* for the respective sub-sequences of X and Y $[X(t_1), ...X(t_i)]$ and $[Y(t_1), ...Y(t_j)]$. The matrix is of size $(n+1, m+1)$ and initialized to zero. The $S_{i,j}$ is then iteratively computed using the following algorithm:

- for $i \in [1, ..., n]$
  - for $j \in [max(1, i-\delta), ..., min(m, i+\delta)]$
    - if $match(X(t_i), Y(t_j))$
      - $S_{i,j} = S_{i-1,j-1} + 1$
    - else
      - $S_{i,j} = max(S_{i-1,j}, S_{i,j-1})$

The maximum value of the matrix is the *LCS*. To be comparable, independently of the indicator respective lengths, the associated similarity measure $LCSS_{\delta,match}(X,Y)$ and distance $DLCS_{\delta,match}(X,Y)$ are typically derived as (*11*)

$$LCSS_{\delta,match}(X,Y) = \frac{LCS_{\delta,match}(X,Y)}{min(n,m)} \tag{1}$$

$$DLCS_{\delta,match}(X,Y) = 1 - LCSS_{\delta,match}(X,Y) \tag{2}$$

$$\tag{3}$$

The parameter δ was introduced in (*11*) to control how far in time elements of the two series can be matched. This is not suited for series that have different lengths, which is not tested in (*11*). As an example (plotted in FIGURE 1), $LCSS_{4,d_{0.1}}([0,1,...19],[10,11,...19]) = 0$ (no similarity) while $[10,11,...19]$ is an exact sub-sequence of $[0,1,...19]$. A solution is to use a simpler version of LCS without δ (which is equivalent to choosing $\delta = +\infty$): $LCSS_{+\infty,d_{0.1}}([0,1,...19],[10,11,...19]) =$ 1 (maximum similarity). This causes other issues as it allows any value to match any other value irrespective of the rate of change in the series (however, the order in the series is always respected). Take for example the series $X = [0,1,...,19]$ and $Y = [0,2,...,18]$ (plotted in FIGURE 1): Y increases at twice the rate of X (with a step of 2 instead of 1), but is still a sub-sequence of X. If for a given application series evolving at different rates of change are considered dissimilar, computing *LCSS* without δ is inappropriate as this example shows: $LCSS_{+\infty,d_{0.1}}(X,Y) = 1$, while $LCSS_{1,d_{0.1}}(X,Y) = 0.2$. Two other examples for two safety indicators, distance and TTC, are showed in FIGURE 3 and illustrate how similarity is over-estimated by the traditional *LCS* computation, while a finite δ does not allow to compute the similarity of the series because they are most similar parts must be aligned.

It follows that the existing formulations of the longest common sub-sequence, with or without δ, are insufficient to measure the similarity of series if the series are simply shifted with respect to each other or if series with different rates of change should be considered different. That is why a new similarity measure is introduced that finds the best alignment of two series while taking into account a finite δ, allowing to take into account the rates of change. The length *ALCS* of the aligned longest common sub-sequence is computed by simply shifting the two series with respect to each other, i.e. by adding an integer parameter $shift$ to the *LCS* computation (replacing the condition $|n-m| \leq \delta$ by $|n-shift-m| \leq \delta$) and taking the maximum *LCS* for all possible $shift$ values. The corresponding aligned similarity measure *ALCSS* and distance *DALCS* are defined accordingly.

Another benefit is to use the longest common sub-sequence itself. The indices corresponding to the elements of the series that are matched to obtain the longest common sub-sequence are obtained by "decoding" the process of the computation of the *LCS*. For example, the longest common sub-sequence of series $X = [1,3,5,6,7]$ and $Y = [1,2,3,4,6,7,8]$, using $d_{0.1}$ and finite δ, are respectively $[0,1,3,4]$ and $[0,2,4,5]$ meaning that the element in position 0 of X matches element 0 of Y, element 1 in X matches element 2 of Y, etc. From these indices can be computed the average difference of the corresponding indices which corresponds to the "optimal" alignment of one series with respect to the other. The alignment corresponding to the *ALCS* is obtained by applying the $shift$ corresponding to the maximum *LCS* to the optimal alignment of the longest common sub-sequence indices. This is very useful to visualize the data and validate the similarities: FIGURE 2 shows the alignment obtained for two TTC indicators considered completely similar if aligned and

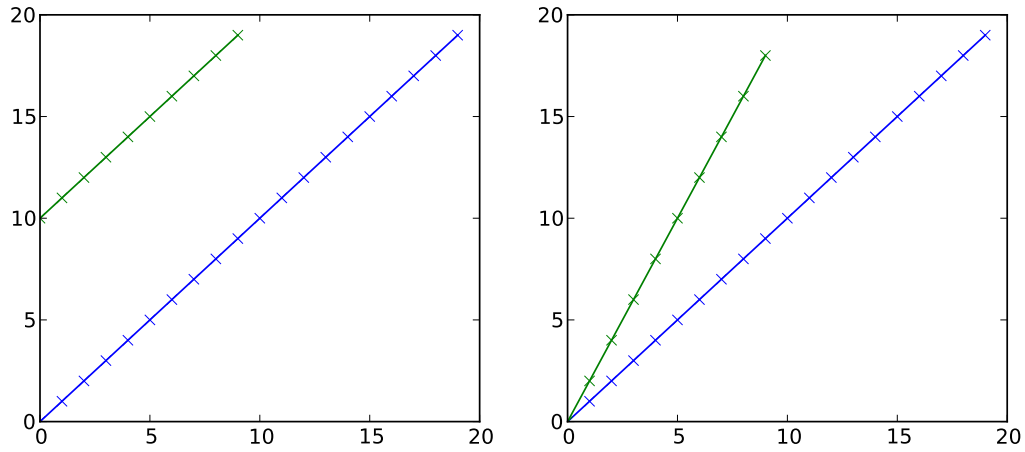


FIGURE 1 Examples of simple series that illustrate the advantages of using a finite δ and aligned longest common sub-sequence. The series in each plot have maximum similarity if using $\delta = +\infty$. This is desired in the plot on the left since it is an exact sub-sequence, but not on the right if the rate of change is taken into account.

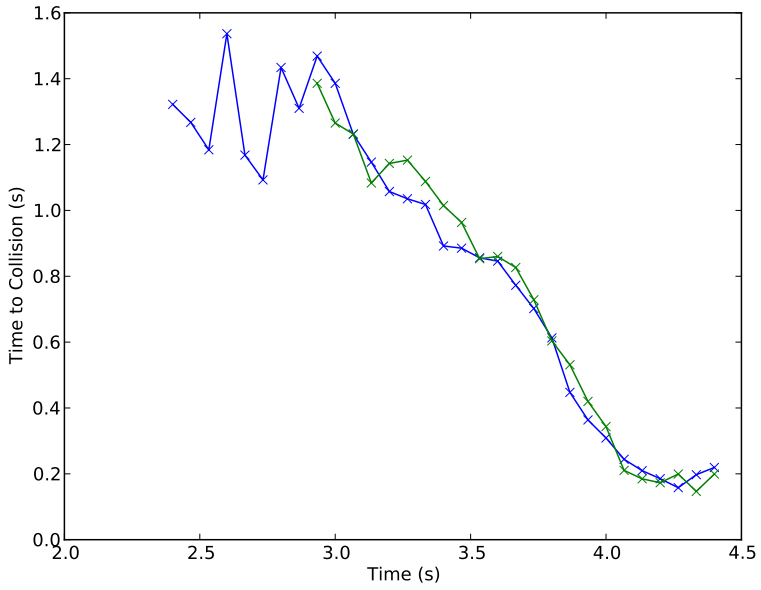
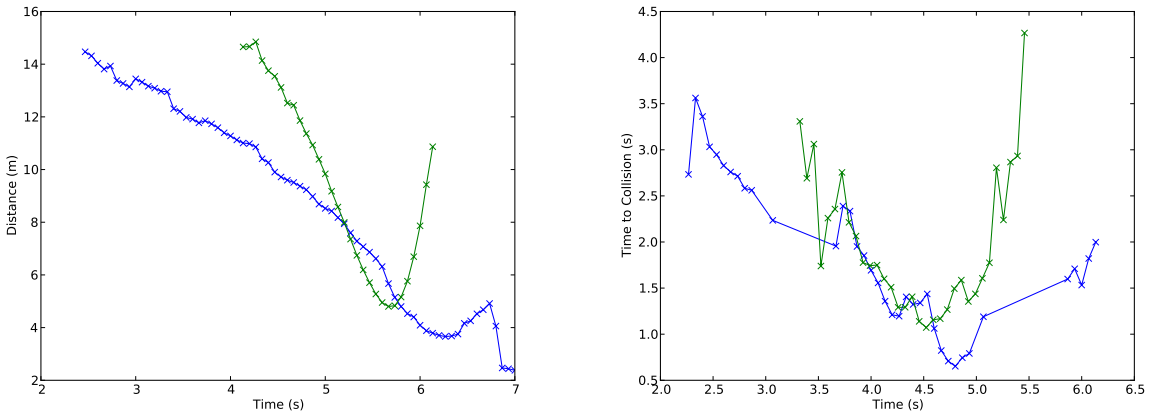


FIGURE 2 Example of alignment of two very similar TTC indicators ($LCSS_{2,d_{0.2\ s}} = 0.2$ and $ALCSS_{2,d_{0.2\ s}} = 1$).

barely similar otherwise ($\delta = 2$ and $d_{0.2\ s}$). The distance and TTC indicators are also aligned in FIGURE 3: these examples of safety indicator profiles should not be considered similar, at least

not to the degree implied by the *LCSS* with infinite δ.



	Distance	TTC
$LCSS_{+\infty}$	0.87	0.64
$LCSS_2$	0.35	0.12
$ALCSS_2$	0.42	0.42

FIGURE 3 Examples of pairs of profiles for the interaction distance and TTC indicators that are more similar using *LCSS* with infinite δ than using *ALCS* and a finite δ. The matching function used is d_ε with $\varepsilon = 1$ m for the distance indicator and $\varepsilon = 0.2$ s for the TTC indicator. The series are aligned according to the aligned longest common sub-sequence.

Clustering Method

The primary goal of this work is to compare road user interactions, characterized by a set of continuous safety indicators. Choosing a data representation and a similarity method rules out some clustering methods. The choice of keeping the indicators in their original shape with variable lengths rules out for example classical clustering algorithms such as k-means since the concept of a centroid is not defined. All the clustering algorithms that operate on a similarity matrix could be used. Several were investigated: spectral clustering was in particular tested at length and used in a first version of this work (*35*). The method is fast and takes as only input the predetermined number of groups. Finding the number of clusters by trial and error proved to be a challenge, and the resulting clusters were not always easy to interpret.

The algorithm used for the results presented in this paper is a slight variation of the algorithm previously developed to cluster motion patterns (*17*). This type of algorithm trades the parameter of the number of clusters for a maximum distance or minimum similarity between instances of the same cluster: when a new instance is to different from the existing clusters, a new one is created for it. The other idea is to use the original data as representatives, or prototypes, for each cluster. That provides a visual and more interpretable representation of each cluster. The last idea is to favour "long" instances, in this case indicators with long time periods of observation. This is done in two ways: first by sorting the indicators according to their length, to start considering first the longest indicators, and second by keeping the longer prototype indicator when two cluster are merged. This solves partially the problem of dependency of the results to the algorithm

initialization which is a well-known limitation of many clustering algorithms such as k-means (initialization of the cluster centroids) and of the one proposed in (*17*). The algorithm parameter is therefore the minimum similarity for two indicators to be in the same cluster: when learning prototypes, an indicator will be added as a new prototype if its maximum similarity to all existing prototypes is lower than the parameter.

EXPERIMENTAL RESULTS

The proposed method to cluster interactions and their safety indicators is tested on a large dataset of 295 traffic videos of collisions and conflicts between motor vehicles collected at an intersection in Kentucky. This unique dataset has already been used in past studies (*2*), most notably in the first paper that compared the characteristics of interactions with and without a collision (*9*). The definition of conflicts used by the people who collected and sorted the data is unknown: a visual review confirms that most match the accepted definition, but will be referred to as interactions without a collision.

As shown in (*2*), an interaction can be well described by several symmetrical indicators based on speed and positions independently of the road user absolute positions. The indicators characterize one road user's motion with respect to the other, as if it was stationary. These indicators based on positions and speed are: the distance of the road users' centroids, the minimum distance separating the road users (from the feature-based tracking algorithm), the speed differential (the norm of the velocity difference), the angle of the velocities, and the collision course angle (the angle between the velocity difference and the vector the links the road users' centroids). To these indicators are added two safety indicators, TTC and probability of collision, calculated in (*2*) using motion prediction methods based on prototype trajectories representing the main motion patterns.

Indicator	Threshold ε	Minimum Clustering Similarity	Number of Clusters
Distance (Dist)	1 m	0.3	6
Speed differential (SD)	1.5 m/s	0.4	4
Velocity angle (VA)	0.15 rad	0.4	4
Time to collision (TTC)	0.2 s	0.3	4
Probability of Collision (PoC)	0.1	0.5	6

TABLE 1 Thresholds ε for d_ε used in the computation of the aligned normalized similarity *ALCSS* with $\delta = 2$, with the minimum similarity used for clustering and the resulting number of clusters.

The choice is made for this study to cluster the interactions based on each indicator separately, for the following ones: distance, speed differential, velocity angle, TTC and probability of collision. For each indicator, a threshold is chosen by trial and error to match the profiles using the aligned normalized similarity *ALCSS* with $\delta = 2$. The matching function is d_ε with the thresholds ε listed in TABLE 1. An additional criteria is added to remove very short indicators that do not contain much information (if not favouring longer indicators in the clustering algorithm, the shortest indicators would tend to be the most similar to the others as they can easily match at least some sub-sequence of a long indicator). The minimum length is 10 frames, i.e. 0.67 s, and

actually applies only to safety indicators since they may not be computed for all instants. The others can be computed as long as the two road users co-exist in the scene. The software code used to compute the similarities, the clustering algorithm and the results presented in this paper are available in the open source Traffic Intelligence project (*6*) and on the page dedicated to this paper (`http://nicolas.saunier.confins.net/data/saunier14trb.html`).

The choice is also made to not display and analyze clusters with too few indicators. The minimum number in the following results is 5 instances, including the prototype. Different minimum similarities for clustering were tested by trial and error for the different indicators and are listed in TABLE 1. For all figures from 4 to 8, each cluster prototype is plotted using dots. Interactions with and without a collision are displayed respectively in red and blue. The numbers beside each cluster number are in order: the percentage of collisions, the number of collisions and the number of indicators in the cluster.

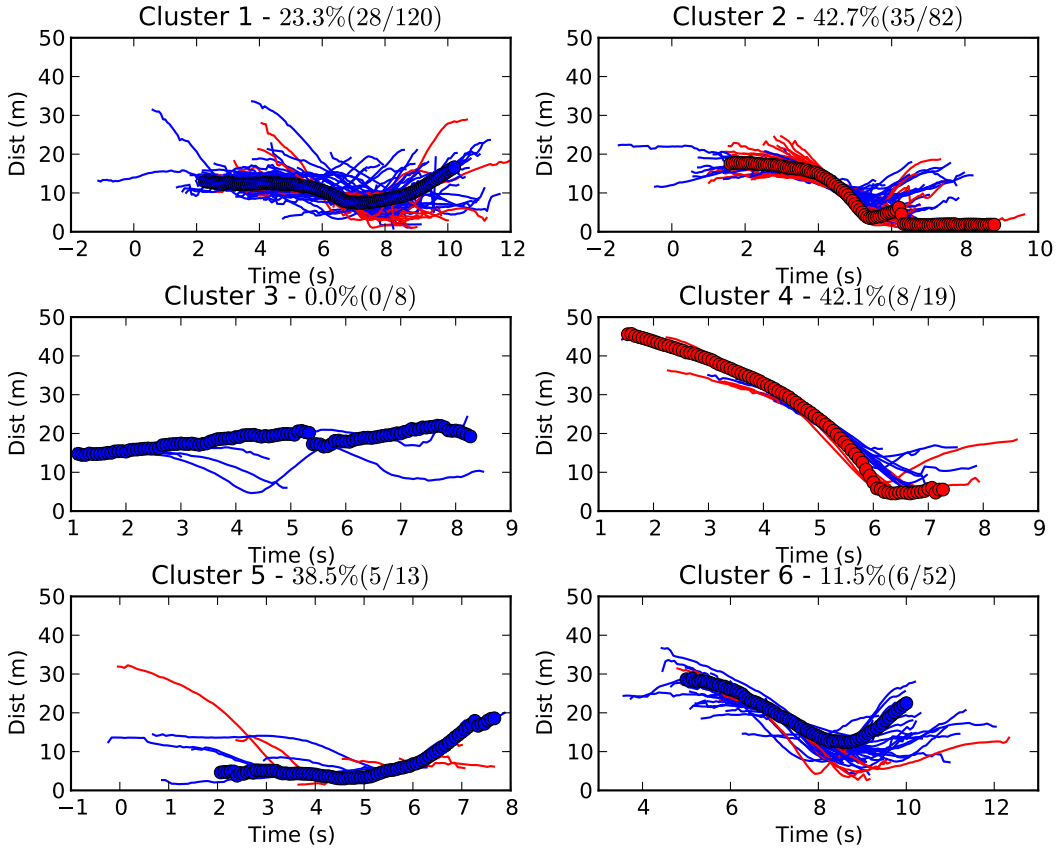


FIGURE 4 Clusters of the distance indicators.

The 6 clusters of distance indicators are plotted in FIGURE 4. There are quite different profiles, from almost flat in cluster 3 to increasing in cluster 5 to decreasing then flat (clusters 2 and 4) or increasing again (clusters 1 and 6). These clusters correspond to varying proportions of collisions: the clusters 2 and 4 contain the most collisions and have therefore expected shapes where the distance remains 0 or close to 0 after the collision. It is however remarkable that a

majority of the interactions in these clusters do not end up in a collision. Clusters 1 and 6 seem to correspond to some sort of evasive action since the distance decreases (the road users are on a collision course), then increases again once the road users start reacting to avoid the collision. The collisions in cluster 5 may correspond to situations where the road users continue moving after the shock. The rate of change differs considerably between the clusters.

The 4 clusters of SD indicators are plotted in FIGURE 5. The shapes are quite distinctive and seem relatively homogeneous for each cluster. There is a pattern relating the proportion of collisions to the highest speed differential: the higher the proportion, the higher the maximum speed differential. This is related to attempts by road users to avoid the collision, which are stronger in collisions. The shape of cluster 1 is particularly striking and could be related to rear end or parallel interactions at similar velocities, followed by a road user turning or changing lane, which puts the road users on a collision course, followed by a return to the initial conditions or more evasive actions with higher speed differential.

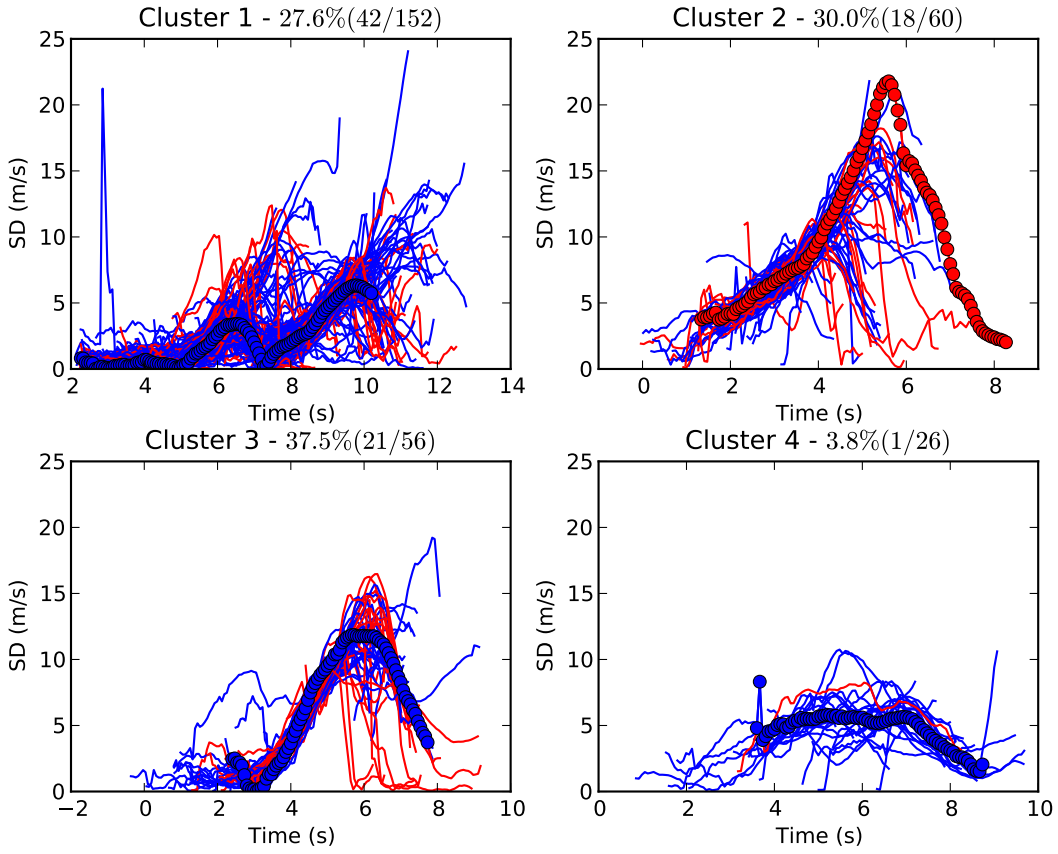


FIGURE 5 Clusters of the SD indicators.

The 4 clusters of VA indicators are plotted in FIGURE 6. The VA indicator is very useful to identify interaction categories, e.g. rear-end, side, head-on, etc., which does change with each instant and is typically not recorded in collision reports. Interactions in cluster 4 are thus side interactions, which evolve as the road users try to avoid each other. Cluster 1 contains even more

collisions and corresponds to rear-end and parallel interactions where the VA increases (some attempt at turning or changing lane) then comes back to 0. The clusters 2 and 3 are more difficult to distinguish. The two must contain situations that start as parallel or rear-end interactions, but evolve into side interactions, with different angles. One cannot miss in any case some important differences in profiles, especially at the beginning in clusters 1 and 2. Trying to obtain more clusters may yield a finer understanding of these clusters.

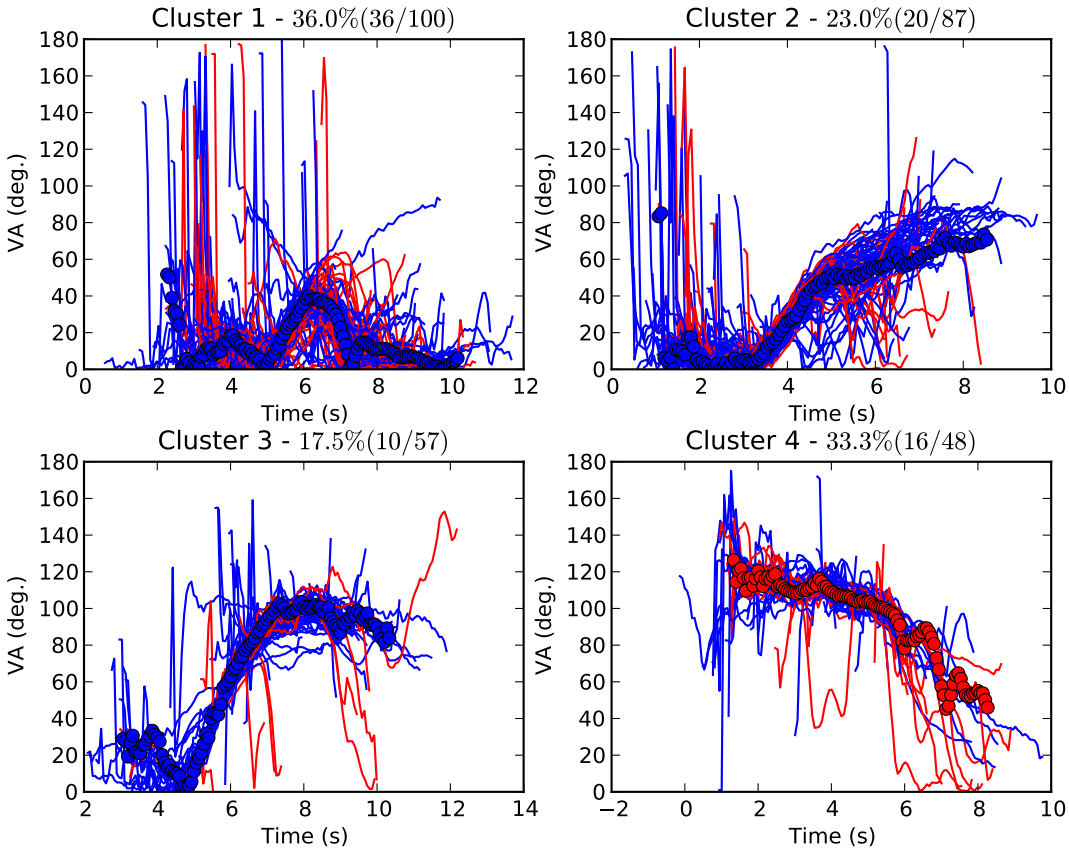


FIGURE 6 Clusters of the VA indicators.

The 4 clusters of TTC indicators are plotted in FIGURE 7. It must be noted that there are only 247 interactions for which TTC can be computed for at least 10 frames. One cannot miss that the TTC indicators are much noisier than the previous indicators. This is related to the quality of the data and the more complex process of computing TTC. Although the motion patterns allow to compute the TTC at more instants, more trajectory prototypes could have made the measures smoother. Most TTC profiles decrease with time as they are expected for collisions and conflicts. The clusters 1 and 2 are very interesting because they look similar at first sight. But the proportion of collisions in cluster 2 is consistent with the profile of its prototype indicator which falls at a seemingly constant rate as a function of time and reaches almost 0 s. On the contrary, there are few TTC measures below 0.5 s or even 1 s in cluster 1. There is more variability in the rate of decrease at the beginning and most profiles increase again after reaching their minimum, which

is consistent with a high proportion of interactions without a collision. Cluster 3 contains mostly collisions, with a higher rate of decrease than cluster 2 which explains why they are in different clusters. Finally, cluster 4 contains only one collision and has fairly constant, noisy, TTC values above 1.5 s.

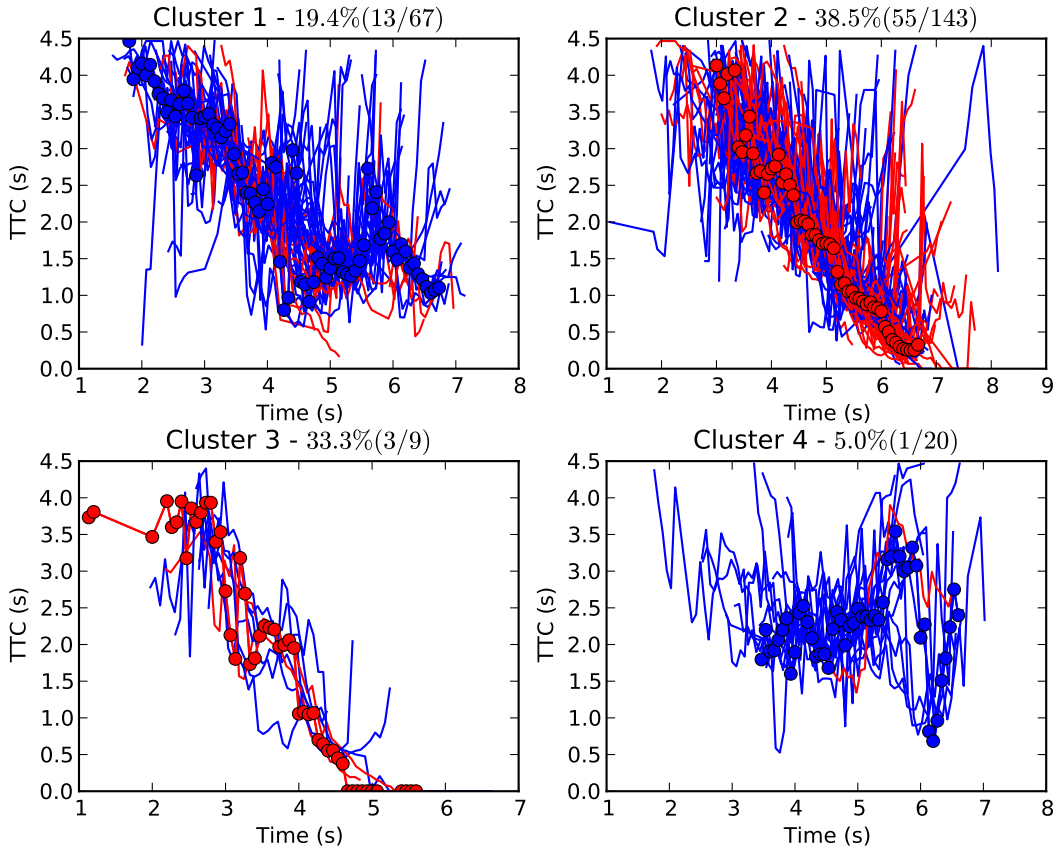


FIGURE 7 Clusters of the TTC indicators.

The 6 clusters of PoC indicators are plotted in FIGURE 8. It must be noted that there are only 260 interactions for which PoC can be computed for at least 10 frames. PoC is also a noisy indicator, depending as TTC on motion prediction methods and the existence of potential collision points. There are two main clusters, 1 and 2, and 4 smaller ones. If ranking the cluster from their maximum PoC, it goes from cluster 3, to cluster 5, then 4 and 6, and finally 2: the first 4 have few collisions, while the last one, cluster 2 has the highest proportion and the highest maximum values reaching 0.8, which is consistent. On the other hand, cluster 1 is more difficult to interpret: it is the largest cluster, contains mostly interactions without a collision, and seems to have two peaks. Whether this is related to noisier interactions without a collision or an actual variation of PoC is unclear.

To sum up the observations, the methods could produce varying numbers of clusters for each indicator that can be interpreted. There are two main results for all indicators. First, there are clusters with very few collisions, e.g. cluster 4 for the SD indicator and cluster 4 for the TTC

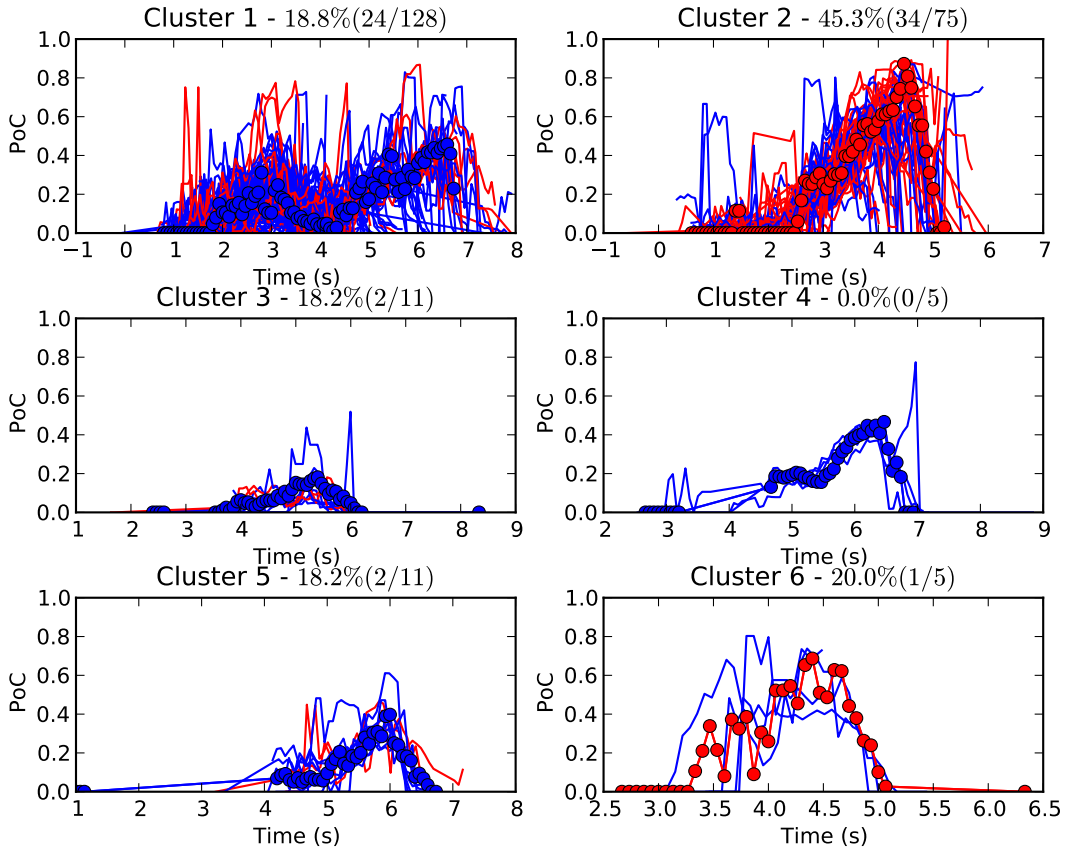


FIGURE 8 Clusters of the PoC indicators.

indicator, which, as was noticed in the previous study (*9*), seems to indicate that some interactions without a collision are not similar to any collisions. This suggests therefore that the factors that are associated with these interactions are different from the ones associated with collisions. Second, it is also clear that even in the clusters with the highest share of collisions (45.3 % for cluster 2 of the PoC indicator), there is always a majority of interactions without a collision that have similar processes to the collisions and are therefore good candidate predictors of these collisions. Finally, there is also a clear trade-off between having few clusters with some degree of variability, which can be seen in several clusters such as cluster 1 for the SD indicator, and more numerous and more homogeneous clusters. This choice is up to the analyst and highlights the flexibility of the method as an exploratory tool.

CONCLUSION

This paper has introduced a new similarity measure built upon the longest common sub-sequence that is sensitive to the rate of change of time series and is shown to be adapted to the clustering of several interaction indicators, including safety indicators such as TTC and the probability of collision. The number of resulting clusters is relatively small and can be easily interpreted in most cases. The results yield further credibility to the main hypothesis of surrogate safety analysis that

some interactions without a collision have similar processes as collisions and could be used as predictors. It also strengthens the observations made in (*9*) that not all interactions should be used for surrogate safety analysis, as can be seen for almost each indicator. Another contribution of this work is to release all the necessary code and data samples to allow true reproducibility of the presented work.

There is considerable room for further research. The main goal is to cluster interactions with and without a collision based on all, or at least several, indicators simultaneously, i.e. by evaluating the similarity of interactions at a given instants through the similarity of all its indicators at this instant. It is hoped that very strong similarities can thus be identified. Finally, this method can be applied to other time series in transportation, especially in safety, such as the large datasets produced by current naturalistic driving studies.

ACKNOWLEDGEMENT

The authors wish to acknowledge the financial support of the Natural Sciences and Research Council of Canada (NSERC). They also wish to thank Zu Kim of California PATH and Ann Stansel of the Kentucky Transportation Cabinet for providing the video dataset.

REFERENCES

[1] Ismail, K., *Application of computer vision techniques for automated road safety analysis and traffic data collection*. Ph.D. thesis, University of British Columbia, 2010.

[2] Saunier, N., T. Sayed, and K. Ismail, Large Scale Automated Analysis of Vehicle Interactions and Collisions. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2147, 2010, pp. 42–50, presented at the 2010 Transportation Research Board Annual Meeting.

[3] Bagdadi, O. and A. Várhelyi, Jerky driving-An indicator of accident proneness? *Accident Analysis & Prevention*, Vol. 43, No. 4, 2011, pp. 1359–1363.

[4] Hallmark, S., D. Mce, K. M. Bauer, J. M. Hutton, G. A. Davis, J. Hourdos, I. Chatterjee, T. Victor, J. Bärgman, M. Dozza, H. Rootzén, J. Lee, C. Ahlström, O. Bagdadi, J. Engström, D. Zholud, and M. Ljung-Aust, *Initial Analyses from the SHRP 2 Naturalistic Driving Study: Addressing Driver Performance and Behavior in Traffic Safety*. Transportation Research Board, 2013.

[5] Saunier, N. and T. Sayed, A feature-based tracking algorithm for vehicles in intersections. In *Canadian Conference on Computer and Robot Vision*, IEEE, Québec, 2006.

[6] Saunier, N., *Traffic Intelligence*. `https://bitbucket.org/Nicolas/trafficintelligence`, 2011-2013, software under the open source MIT License.

[7] *The Swedish Traffic Conflict Technique*. Sweden, 2005.

[8] Laureshyn, A., Å. Svensson, and C. Hydén, Evaluation of traffic safety, based on micro-level behavioural data: Theoretical framework and first implementation. *Accident Analysis & Prevention*, Vol. 42, No. 6, 2010, pp. 1637–1646.

[9] Saunier, N., N. Mourji, and B. Agard, Investigating Collision Factors by Mining Microscopic Data of Vehicle Conflicts and Collisions. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2237, 2011, pp. 41–50, presented at the 2011 Transportation Research Board Annual Meeting.

[10] Svensson, A. and C. Hydén, Estimating the severity of safety related behaviour. *Accident Analysis & Prevention*, Vol. 38, No. 2, 2006, pp. 379–385.

[11] Vlachos, M., G. Kollios, and D. Gunopulos, Elastic Translation Invariant Matching of Trajectories. *Machine Learning*, Vol. 58, No. 2-3, 2005, pp. 301–334.

[12] Svensson, A., *A Method for Analyzing the Traffic Process in a Safety Perspective*. Ph.D. thesis, University of Lund, 1998, bulletin 166.

[13] Archer, J., *Methods for the Assessment and Prediction of Traffic Safety at Urban Intersections and their Application in Micro-simulation Modelling*. Academic thesis, Royal Institute of Technology, Stockholm, Sweden, 2004.

[14] Laureshyn, A., *Application of automated video analysis to road user behaviour*. Ph.D. thesis, Lund University, 2010.

[15] Tarko, A., G. A. Davis, N. Saunier, T. Sayed, and S. Washington, *Surrogate Measures of Safety*. White paper, ANB20(3) Subcommittee on Surrogate Measures of Safety, 2009.

[16] Amundsen, F. and C. Hydén (eds.), *Proceedings of the first workshop on traffic conflicts*, Institute of Transport Economics, Oslo, Norway, 1977.

[17] Saunier, N., T. Sayed, and C. Lim, Probabilistic Collision Prediction for Vision-Based Automated Road Safety Analysis. In *The 10^{th} International IEEE Conference on Intelligent Transportation Systems*, IEEE, Seattle, 2007, pp. 872–878.

[18] Mohamed, M. G. and N. Saunier, Motion Prediction Methods for Surrogate Safety Analysis. In *Transportation Research Board Annual Meeting Compendium of Papers*, 2013, 13-4647. Accepted for publication in Transportation Research Record: Journal of the Transportation Research Board.

[19] 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.

[20] Gettman, D., L. Pu, T. Sayed, and S. Shelby, *Surrogate Safety Assessment Model and Validation: Final Report*. FHWA, 2008.

[21] Minderhoud, M. M. and P. H. Bovy, Extended time-to-collision measures for road traffic safety assessment. *Accident Analysis & Prevention*, Vol. 33, No. 1, 2001, pp. 89–97.

[22] Parkhurst, D., *Using Digital Video Analysis to Monitor Driver Behavior at Intersections*. Center for Transportation Research and Education (CTRE), Iowa State University, 2006.

[23] Laureshyn, A., K. Åström, and K. Brundell-Freij, From speed profile data to analysis of behaviour. *IATSS Research*, Vol. 33, No. 2, 2009, pp. 88–98.

[24] 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.

[25] Alpaydin, E., *Introduction to Machine Learning, second edition*. The MIT Press, second edition ed., 2010.

[26] Vlachos, M., G. Kollios, and D. Gunopulos, Discovering Similar Multidimensional Trajectories. In *Proc. of 18th International Conference on Data Engineering (ICDE)*, San Jose, CA, 2002, pp. 673–684.

[27] Liao, T. W., Clustering of time series data: a survey. *Pattern Recognition*, Vol. 38, No. 11, 2005, pp. 1857–1874.

[28] Vlachos, M., J. Lin, E. Keogh, and D. Gunopulos, A Wavelet-Based Anytime Algorithm for K-Means Clustering of Time Series. In *Proc. Workshop on Clustering High Dimensionality Data and Its Applications*, 2003, pp. 23–30.

[29] Morris, B. and M. Trivedi, Learning trajectory patterns by clustering: Experimental studies and comparative evaluation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 312–319.

[30] Piciarelli, C. and G. Foresti, On-line trajectory clustering for anomalous events detection. *Pattern Recognition Letters*, Vol. 27, No. 15, 2006, pp. 1835–1842.

[31] Bennewitz, M., W. Burgard, G. Cielniak, and S. Thrun, Learning Motion Patterns of People for Compliant Robot Motion. *The International Journal of Robotics Research*, Vol. 24, No. 1, 2005, pp. 31–48.

[32] 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.

[33] Morris, B. and M. Trivedi, Trajectory Learning for Activity Understanding: Unsupervised, Multilevel, and Long-Term Adaptive Approach. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, Vol. 33, No. 11, 2011, pp. 2287–2301.

[34] Zelnik-Manor, L. and P. Perona, Self-Tuning Spectral Clustering. In *Advances in Neural Information Processing Systems 17* (L. K. Saul, Y. Weiss, and L. Bottou, eds.), MIT Press, Cambridge, MA, 2005, pp. 1601–1608.

[35] Mohamed, M. G. and N. Saunier, Classifying Profiles of Surrogate Safety Measures to Understand Collision Processes. In *Canadian Multidisciplinary Road Safety Conference*, Montreal, 2013.