Visual Analysis of Car Fleet Trajectories to Find Representative Routes for Automotive Research

David Spretke*, Manuel Stein*, Lyubka Sharalieva*, Alexander Warta†, Valentin Licht†, Tobias Schreck* and Daniel A. Keim*

*University of Konstanz
Konstanz, Germany
Email: {david.spretke, manuel.stein, lyubka.sharalieva, tobias.schreck, daniel.keim}@uni-konstanz.de

†Robert Bosch GmbH
Stuttgart, Germany
Email: {alexander.warta, valentin.licht}@de.bosch.com

Abstract—Testing is an important and widespread practice in the development of automotive components. For the design of test methods two types of input data are often considered: (1) load data gathered from real life vehicle fleets, and (2) information of the driving routes based on road features. The development of new technologies is though complicated not only by the need to join those two data sources, but also by the too limited knowledge of the parameters and their useful combinations. As a result, information about representative driving profiles is needed. To address these problems we present a visual analytics approach for analyzing multivariate trajectories as a combination of vehicle’s location and road elevation data. Our system combines trajectory clustering, interval-based user-driven trip segmentation, and frequent sequences analysis, supported by contingency table and interval-based Parallel Coordinates visualization and enables the expert user to find representative driving profiles for the definition of very compact test courses.

Keywords—Visual Analytics, Automotive Research, Trajectory Analysis and Visualization

I. INTRODUCTION

The increasing demand on the automotive industry for effective and efficient vehicle solutions leads to a growing need for data analysis activities on car fleet test data. In the advent of ubiquitous computing, the acquisition of the location of moving objects, e.g., vehicle’s locations from global positioning system (GPS) and radio-frequency identification (RFID) can make this task easier. Tremendous amounts of data about the test car fleet on the road networks is being collected, which is even more difficult to be analyzed due to their huge size. The importance of visual analytics of trajectory data, e.g., vehicles’ trips, is thus widely recognized. Handling this complicated task benefit also other research areas beside the automotive industry. Some broad research fields are the analysis of driving information for traffic estimation [1], [3], movement behavior, e.g. finding frequent driving routes [18], [17] or manoeuvr analysis and microscopic traffic pattern finding [14], [13], [11]. Further, the analysis is not restricted to the automotive data, but is also applied to behavioral ecology as animal [26], [22] or vessels movement [28], etc.

As we can see, the challenge of the analysis of trajectory data applies to more than one research area. In the automotive industry, where huge amounts of GPS data are collected, one of the main applications is the development of automotive components, where testing plays an important role. The testing builds on already known features of the driving profile, which include not only the most frequent driving locations, but also vehicle data, e.g., engine temperature or torque, and information about the road’s elevation profile. All of the mentioned data introduces a combination of information sources, which need to be automatically and visually analyzed in order to gain significant testing information.

Despite the extensive research on trajectory data, current work on trajectory analysis neglects an important aspect that lies in between the actual moving data and the road features, i.e. the elevation profile, with its absolute height and slope values. This paper proposes a visual analytics approach for analyzing such multivariate trajectories as a combination of vehicle’s location and road elevation data.

We claim the following contributions: 1) a suitable and very compact representation of the elevation profile of trajectories; 2) a visual analytics system for analyzing car fleet data that combines trajectory clustering with feature and location based self-organizing maps visualization, geographic map visualization, interval-based user-driven trip segmentation, and frequent sequences analysis, supported by contingency table and interval-based Parallel Coordinates visualization; and 3) two case studies and a user evaluation with qualitative comparison of the implemented methods based on functionality, usability and achieved results, conducted with domain users.
II. RELATED WORK

We first discuss related work based on the general analysis of nowadays data mining methods, followed by specific work, in Section II-A. Section II-B introduces to related approaches for visual analysis of trajectory and movement data and positions our approach within the aforementioned works.

A. Analysis Based on Data Mining Methods

Important data analysis methods in this field include the abstraction, classification, clustering, correlation or segmentation of trajectories. Today, many applications for trajectory-based data analysis have been identified, including studying of traffic data [27].

Semantic segmentation is a related data mining approach, which assigns objects to one of several semantic classes [9]. This approach is widely used in the recognition of objects on pictures, where the objects are represented by pixels with specific color [4], [23]. In the trajectory analysis, semantic segmentation is mostly based on grouping of GPS points, which are aggregated, e.g., using their spatio-temporal morphology (topology) and stop-overs to define action classes [13]. Sequential pattern mining beside as another related data mining method, which assigns objects to one of several semantic classes [9]. This approach is widely used in the recognition of objects on pictures, where the objects are represented by pixels with specific color [4], [23]. In the trajectory analysis, semantic segmentation is mostly based on grouping of GPS points, which are aggregated, e.g., using their spatio-temporal morphology (topology) and stop-overs to define action classes [13]. Sequential pattern mining beside.

B. Visual Analysis of Trajectory and Movement Data

With the rapid development of location acquisition methods, the focus on trajectory data increases and thus the need for methods to visually analyze this data. The work most related to the subject of our approach is introduced by Fazekas et al. [13], where the analysis on route data of truck fleets is performed in two steps - segmentation of routes into meaningful route sections and automatic rendering of the sections into activity classes. Between the work of [13] and the innovative method introduced in our approach there are some basic resemblances. For example, both methods work with real life vehicle data, provided by recorded trajectory and speed data, as well as transportation specific and general road data. Further, analysis of the vehicle trajectories is performed after meaningful sampling and creation of segment sequences. Nevertheless, there are some differences between the methods worth mentioning. Fazekas et al. [13] concentrate on low-speed and stopping behavior of vehicles, expanding its previous work on driving safety evaluation and breaking data [12].

Figure 1. Analysis steps of the ST-Matching algorithm: In (a), the candidates for the matching are retrieved, filtered and rated. (b) Afterwards, the path between each combination of candidate points of two tracks is compared to the air line between the tracks and also rated accordingly.

Trajectory data analysis can be applied not only to vehicle data, but to other data types, e.g., animal movement data. Animal movement analysis is performed by, e.g. Spretke et al. [26], where a complex visualization system supports the analysis. As stated in their previous paper [21], there are three main challenges, that need to be considered when implementing a system for animal movement analysis: the complexity of the data, the integration of heterogenous data sources and the need to flexible set and explore parameters. Another approach for visual analysis of moving objects can be found in [28], where a database of vessel traffic data is considered.

III. DATA AND ATTRIBUTES

We consider a car fleet data set collected in research by Robert Bosch GmbH which includes 19 Mio. GPS tracks (with latitude, longitude and time stamp) in Europe covering 1.5 years. The tracks represent a heavy duty truck fleet and include 269 vehicles, which drove 568.869 km in total. Initially, the GPS tracks contained 52 attributes, from which the most valuable is the vehicle’s mileage at a track with a specific time stamp. The mean distance between the GPS positions lie between 5 and 15 minutes. Having this data our goal was to find representative driving profiles for the definition of very compact test courses.

The GPS tracks were preprocessed using a map-matching algorithm, which aims to find the real driving route between two GPS tracks by using the track’s location and time stamp. For matching GPS tracks on a map, only few algorithms consider low-sampled data. The Spatio-Temporal (ST) Map-Matching Algorithm of Lou [19] is one of them. The algorithm is built on a probability estimation of the possible routes, where the analysis starts with a candidate preparation. In this step the possible nodes on a map within a user-defined radius and limited candidate number are found and rated based on their distance to the track, as shown in Figure 1(a). Further, spatial analysis is performed, which considers the likelihood that the route between two GPS tracks follows the shortest path between the candidate points (see Figure 1(b)). In a last step, the speed limitations of the road are considered. In a resulting graph, all probabilities...
are multiplied and thus the most probable non-intermittent route is chosen as a resulting trip.

The ST-Matching algorithm [19] returns trips of moving and non-moving segments. GPS tracks with time difference less than 60 minutes between them are combined in sequences, since after such a long time interval one can assume that a new trip has started. These track sequences are given as input for the algorithm. As a result the algorithm produces trips - sequences of road elements from the map graph, representing the edges between nodes - which are built by starting, ending and inner nodes including the node position in latitude and longitude format and information about the road, e.g., road’s elevation, maximum speed, etc. The inner nodes are placed in a manner that optimally describes the map topology. Further inner nodes are also added where the road height changes.

The most important attributes for the discovery of representative height and street profiles are the road type and the road elevation, which therefore are integrated into a feature vector (FV): id, length, cc_highway, cc_rural, cc_city, up_to_length (uphill slope), down_to_length (downhill slope) and change_to_length (uphill downhill slope changes). The parameters, according to the course composition - cc_highway, cc_rural and cc_city - express the percentage of each road type within the trip. The course composition (CC) is a standard way of characterizing a trip in the automotive research and its three values sum up to 100%, since all three attributes are expressed as percentage of the overall trip length. The last three attributes in the feature vector represent the uphill and downhill slope and the uphill downhill slope changes. Additional attributes, which are not part of the data mining process, were also considered, e.g., the absolute mileage of the highway, the rural and the city parts of each trip, the absolute value for the travelled uphill and downhill height and their difference and the absolute number of changes within each trip. These values were though normalized by the length, so the trips can be made comparable to each other.

IV. VISUAL ANALYTICS OF CAR FLEET TRAJECTORIES

Our system analyzes the car fleet data on two different granularity levels - the first one (MT1) performs a clustering on the trip’s feature vectors, the second one (MT2) handles the road elements and explores their feature vectors. Both methods incorporate different procedure: MT1 starts with intensive visual and interactive user support, then requires the user’s knowledge by the definition of representativeness and the discovery of the profile. In comparison, MT2 implies the user’s domain knowledge in a very early stage during the selection of the boundary values for the segmentation. Here the user is assisted by a visual validation step. Finally, a comparative visual analysis of frequent sequences of elevation events can be made and the representative profile can be chosen. Figure 2 shows a schematic overview of the system and both methods are now described in detail.

A. Trip Clustering and Visualization

This method (MT1) performs an analysis on the trips of the car fleet data (low granularity level) and combines data mining steps with several visualization techniques. The analysis starts by displaying an overview visualization of the database information to the user, as shown in Figure 3. Parallel coordinates [15] display the trip’s feature vectors with the six attributes - relative highway, rural and city mileage, uphill and downhill slope and height changes. The mean database vector is also visualized with a red line. Here, the user can make observations about the overall distribution of the values, the mean vector and the minimal and maximal attribute values and for this tasks she or he is supported by

Figure 2. Schematic view of the system with two different methods. (i) marks parts of the system where the user is able to interact with the tool.

Figure 3. Database overview: Parallel Coordinates and Box Plots display the feature vectors of all trips in the car fleet data.
the yellow vector color, a certain opacity and overlapping, so frequent values may be brought out.

As suitable procedure to find the main car fleet applications in the data we’ve chosen a standard clustering approach - Self-organizing maps [16] (SOM). The SOM method maps high-dimensional feature data to a grid of cluster prototypes, approximately preserving data similarity in the map layout. The techniques has been successfully applied to geospatial data analysis tasks before, e.g., in [2] to analyze time series data in geospatial context.

The standard method in the automotive industry for analyzing car fleet data based on the road type is using empirical statistics for creating mean values or computing attribute distributions. In this way, knowledge about the course composition of the whole car fleet data can be computed. Additional attribute distributions are analyzed simultaneously, but separately from the rest of the data. In our case, clustering shows to be suitable for the automatic analysis of the road information including the elevation profile or even further attributes in the future and may introduce a fast and easy to use and understandable alternative to the standard empirical statistics.

One of the most significant advantages of the SOM technique is the map, produced as a result of the clustering. Every cluster includes trips with similar feature vectors, the clusters though are related to each other and visually ordered in a map, so trends can be easily found. Furthermore, the integration of two visualization techniques - Parallel Coordinates (PC) and geographic map - into the SOM as Small Multiples enhance the user’s understandability of the profile (see Figure 4). Parallel coordinates have shown to be a good and frequently used approach to visualize a big amount of multidimensional data [15]. Further, the mean vector, i.e. the cluster representative, and the cluster coherency can be observed. The representative also plays a big role in the determination of the representative profile, since it is the most representative feature vector of this cluster. A geographic map is included, on which the cluster’s trips are visualized. This is a step towards supporting the user to add an intuitive visualization to the otherwise rather abstract PC representation and values. Here, the user is able to understand the profile of the cluster by analyzing where the trips occurred. As a result, a link can be made effectively between the profile and the actual geographic position.

The clustering is supported by visual highlighting of clusters with special features (e.g. number of objects in the cluster, total or average mileage, customized ranking), where the cluster with the highest value will get the color of the biggest hue. Accordingly, the discovery of a represented cluster based on the SOM coloring is supported. Furthermore, a detailed analysis of the clusters is possible to support the user in selecting a representative profile.

B. Trip Segmentation and Frequent Segment Order

Our second method (MT2) helps to discover a more detailed representative profile by analyzing the road elements of the car fleet data (high granularity level) and combing interval-based user-driven segmentation of car fleet trips and automatic frequent sequences analysis. As mentioned earlier, the chosen data representation incorporates some level of information loss. This origins in the definition of the very aggregated feature vector, including only 3 course composition and 3 elevation description attributes. Within this approach, we aim to discover a more detailed representative profile using a combination of data mining techniques.
As described in the last section, the map matching results in trips, containing moving and non-moving segments, which are build by road elements. While MT1 only analyzes the upper most level of the matched data, the trip level, MT2 now performs a more detailed analysis on the road elements. The different road elements with specific elevation and course composition may be recognized as events. Similar to MT1, the analysis starts with charts, displaying main data base characteristics (see right part of Figure 5(a)). As first and most important, a course composition (CC) is displayed in a bar chart, since every road element is of one of the three road types - highway, rural and city, and the CC is the most important information beside the elevation profile. Additional three charts visualize the same feature vector values of the elevation, but here the road element’s height information is displayed.

By user-defined height attribute classes, a segmentation-based approach splits the trips into smaller driving segments with specific characteristics, in future referred as events or trip segments. Since the segmentation relies on the user knowledge and sufficient boundary choice, a visual validation step is built in, so the resulting segments will be displayed in a contingency table (see Figure 5(a)), which also reminds to the SOM map from MT1. Similarly, the analysis of the segment classes is visually supported by Parallel Coordinates (PC). According to the class distribution, the user decides, whether the segmentation was successful or a boundary reconsideration is needed. This analysis step is relatively scalable, since a contingency table with Parallel Coordinates may display many trip segments. Indeed, this feature is very important, since one can imagine that specific boundary selections may result in a huge amount of segments with short distance.

For the user it could be important to find the exact ordering of specific important elevation and road type events, e.g. it may be important whether the representative trip profiles starts or ends in a city region with steep downhill elevation profile. This task is performed by a sequential pattern mining algorithm. We chose to use the SPAM algorithm [5], as it has been shown to provide robust results, however note we could rely on other methods as well. As a result, a new window opens (Figure 5(b)), which includes the most frequent sequences of events and additional information about their number, length, the number of the used segments and options regarding the input parameters of the SPAM algorithm as minimal support and minimal pattern length. In the left part of the same window are displayed the sequences according to the specified parameter. In Figure 5(b) we see four patterns, visualized by a bit representation and interval-based parallel coordinates. Here, the user plays a big role in the decision of the suitable minimal sequence frequency and the selection of the representative profile.

V. Evaluation

A. User Study

A user evaluation with four domain experts, aged between 31 and 46 years, was conducted with Bosch research to perform a qualitative evaluation of the implemented system. The evaluation was performed on a notebook, connected to a monitor, where the software was started and displayed. All of the invited experts are working in the automotive business, mainly in the reliability testing of hybrid engines,
although they come from different fields of study - Mechanical Engineering, Mechatronics and Information systems. In this working area the experts have between 3 and 19 years working experience and work most of their time on a computer.

The study was built in a qualitative manner, producing a comparison and method description in non-numeric terms. The users were asked to work with both methods (MT1 and MT2), solving a list of tasks, while thinking aloud. Further, they were asked to explain what each task means for the analysis and to what extent they understand the system. In the end of the study, an expert interview was performed and a questionnaire was filled out. Each method evaluation included three main tasks, defined as follows:

1) Introductive tasks about the database, aiming to allow the user to get insight into the represented data and the visualization techniques.
2) Tasks for evaluation of the data analysis techniques and the method’s ability to visually illustrate the results in an easy to understand way.
3) Tasks for discovery of representative profile.

In general, the results of the evaluation are satisfying. All domain experts were very motivated to participate in the study. They performed tasks with both methods and shared their thoughts. One of the users, who took his university degree in the field of Information systems, oriented his attention mainly on the visualization of both methods. Only one of the participants had problems by the analysis and needed intensive task support.

Performing the first task of the first method (MT1) was easy and information was comprehended fast. Analyzing the database information in form of trips feature vectors and box plots appeared to be an easy task, where all users found the most important information without problems. The overall satisfaction from the method continued while performing task 2, where the SOM with the integrated maps was evaluated. Here, the map got particular appreciation, since some of the users identified it as the most important and interesting visualization part. The experts could mentally combine the cluster’s Parallel Coordinates and easily give meaning to the values. The most important task was to find the representative profile in the database. Regarding the definition of representativeness, many users chose the same cluster, the one with the best ranking (color hue) and the best distribution across Germany. The evaluation of the second method (MT2) took other course. Independent of the ordering of the methods, MT2 took more time, more support from the study moderator and was generally difficult to understand. Some of the users described this method as rather abstract and needing a big amount of domain knowledge. As expected, all of the users liked MT1 better and found that they could come to the desired results easier than with MT2. Techniques as Parallel Coordinates and SOM with map view were positively rated and the users could discover a representative profile. About MT2, the users wished for more extensive automatic support by the class boundary selection and better and more intuitive visualization of the frequent segment sequences.

B. Case Studies

1) Case Study 1: Operational Profiles: The car fleet contains different operational profiles of the trucks, which may differ greatly from one another and hence distort the mean feature vector in a certain direction. Finding groups of similar trips and their frequency and mileage is beneficial from the clustering, which produces 9 different clusters with different driving profiles. Two examples of truck applications are shown in Figure 6 - a "city" and a "flat" cluster. These two applications represent the upper and lower extremum value intervals of the last feature vector value - uphill downhill slope changes. The violet cluster represents 10% of the trip sample, but only 1% of its mileage. It contains only short trips with average trip length of 39 km. Additionally, the course composition of this cluster represents only 50% highway, which is the only cluster with such a low highway percentage. Based on its abnormal features, this cluster can be recognized very fast within the SOM view. Analysis on the map further enables the user to specify the trip’s most frequent locations near big German cities, including Berlin, Magdeburg, Essen, Mainz and Ludwigshafen. Since the trucks ship goods from one city to another, this trip cluster may be a result of being stationary while waiting for new goods. Further, it is possible, that these are the main locations of the facilities of the organization, from which this car fleet information was collected.

The yellow cluster contains slightly more trips, but also more mileage and can be easily defined as a "flat" cluster, based on its feature vector and trip location. The cluster represents only the most lower interval of all of the three height profile attributes in the feature vector and the locations, where the trips took place are also with less slope and changes - near the sea in northern Germany and along the Rhine river. Although this cluster mean feature
vector do not overlay with the mean data sample feature vector, this application may be important for specific truck measurements, requiring such terrain.

2) Case Study 2: Finding profiles according to the Frequent Sequence Order: In Figure 5(a) we see the segment class distribution for the attributes uphill and downhill slope after segmentation using MT2. In the cumulative line charts in the attribute’s panel we identify the boundaries for these attributes, which are placed in a manner, that separates the route elements in 50%, 25% and 25% groups, i.e. the lower boundary is set to 0.1% and the upper boundary - to 0.5% for both attributes. This separation is based on the assumption, that road elements with uphill slope less than 0.1% are "flat", if the downhill slope also possesses such low value. The figure shows that the number of segments over each row (i.e., each class for the average uphill slope for the road element building the segments) almost exactly equals the chosen distribution.

Further analysis arose questions about the distribution of the trip segments along the road type and attribute classes. An interesting finding is that the segmentation of the trips with the same attribute boundaries for uphill and downhill slope did not produce the same class distribution. This insight is counter intuitive since one may think that the slope values of both attributes occur with the same frequency. Nevertheless, the contingency table displays the distribution of the trip segments resulting from the segmentation, hence slightly different number of segments in each class is possible.

In a visual comparison between the sequences in Figure 5(b) we see, that all four sequences end with two segments - the first with high downhill and the second, which differs also by its road type, namely city instead of highway, with high uphill slope. The proceeding subsequence differentiates between the patterns, but always starts with changing middle uphill and downhill slope and height changes in the middle class. This may be labelled as "hilly" pattern and is extending from just one segment in the first sequence to five sequential segments. To sum up, all four frequent sequences in this example include at first hilly sub-pattern and end with long steep downhill then uphill city segment. Hence, these patterns may appear in result of driving events, where the truck was driving some time on the highway, then took a highway exit to get to a truck parking lot. A further possibility for these sequences is that trucks drove from the highway road onto a resting place. Interesting analysis steps may be to decrease the frequency to explore whether these sequences will be extended at the end by further city segments, which will be the first case, or by highway segments, which will favor the second case.

VI. CONCLUSION

This paper described a visual analytics system for the analysis of car fleet trajectories. Our system supports the user in finding representative driving profiles based on road and elevation features, and the derivation of representative driving routes. We combine various mostly automatic data mining techniques to analyze a new and only less explored field of study - the elevation profile of trajectories in the field - and support this task due visualization techniques and heuristic functions. The analysis is performed on two granularity levels and thus enables fast and easier representative driving route discovery as well as more detailed sub-trajectory analysis.

Based on the findings of the evaluation, some improvements of the system could be recognized. Regarding MT1, a possible enhancement may be an automatic identification of the SOM parameter (map size, learning rate, neighbourhood size) [6], including more cluster highlighting options or allowing the user to define combinations, e.g., implementing the ranking function more flexibly. Within the second method, MT2, there is an enhancement potential within the analysis. An important information, gained from the user evaluation, is the need of extensive automatic user support by the segmentation and more suitable and more self-explanatory pattern visualization by the event sequences. Further, the user might define profile events not only based on the attribute frequency, but also using different criteria or by drawing or giving an example.

ACKNOWLEDGMENT

This paper is based on the master thesis "Visual analytics of multivariate trajectories for the derivation of representative driving profiles" of Lyubka Sharalieva, University of Konstanz.

REFERENCES


