International Workshop on Visual Analytics (2012)
K. Matkovic and G. Santucci (Editors)

Visual Analysis of Public Transport Vehicle Movement

M. Wörner[1,2] and T. Ertl[2]

[1]GSaME Graduate School for advanced Manufacturing Engineering, University of Stuttgart, Germany
[2]Institute for Visualization and Interactive Systems, University of Stuttgart, Germany

Abstract

Public transport vehicles record detailed operational logs. In major networks, this quickly amounts to massive datasets. Traditionally, these are analysed only using statistical tools, yet they show great potential for the successful application of visual analytics techniques and the integration of different visualizations in order to improve service quality and passenger comfort. We present a visual analytics tool for the analysis of such data and apply it to a real-world dataset of actual public transport vehicle movements.

Categories and Subject Descriptors (according to ACM CCS): I.3.8 [Computer Graphics]: Applications—

1. Introduction

Local public transport is an important element of the infrastructure of major cities. It usually combines buses, trams and local trains to create a dense network of connections between stations distributed across an urban area. Much care is taken to create schedules which result in a good overall supply of transportation services, taking into account the time-dependant demand for certain connection, the capabilities of the infrastructure, and the available resources (vehicles, personnel, funds). Once in effect, schedules need to be verified regularly to make sure they are efficient, effective, and can be maintained in practice. From time to time, schedules will have to be adapted to the changing realities of a living city where both infrastructure and demand will change over the course of the years. In the shorter term, ad hoc schedule changes need to account for accidents, temporary blockages, and temporary changes in demand, such as those caused by major local events. In all these cases, the overall goal is to provide a reasonable quality of service to the customers with the resources and funds available.

In this day and age, traffic networks collect detailed data on their current state. Vehicles automatically record their speed and location and report stopping at stations. Schedules and assignments are available as structured data tables and there is usually a detailed digital map of the stations and possibly the roads or rail tracks connecting them. Vehicle states, sampled every few seconds, quickly accumulate to a considerable amount of data, which is traditionally used for general statistical reports. It is no surprise that Thomas and Kielman list the application to the transportation domain as one of the challenges for visual analytics [TK09] and that it is an active research topic. [vFHHP09] analysed the properties of the public transportation network infrastructures of several major cities. They examined stations and routes but no actual vehicle movements. The authors of [KVJ08] discussed the possibility of tracking vehicles using odometer values without a geometric infrastructure definition by using digital maps and image processing. [SWvdWvW11] aggregated and visualized large numbers of local ship trajectories and [LJF08] created a visual analysis system for analysing vehicle movement on the much larger scale of intercontinental shipping routes. [AA08] presented a method for the visual analysis of a massive number of vehicle trajectories not bound to public transport routes or rail tracks. Flowstrates [BBBL11] is an approach to visualize movements between a number of destinations over time using a matrix-like view. [AA05] explored several methods for preventing overcrowing in time-series graphs with many data sequences.

A transport company running trams and buses in a major city provided us with a real-world dataset. We discussed their current procedures and tools with their domain experts and have started to explore how visual analytics tools and methods might improve upon these and aid in better utilizing this data. As a first analysis topic, we examined vehicle dynamics. Technical means ensure that safety margins are not exceeded, but negative effects on passenger comfort might go unnoticed.

DOI: 10.2312/PE/EuroVAST/EuroVA12/079-083

2. The data

The data we received includes infrastructure data, specifying the location of all stations and the geometry of the rail track network. There are route pattern tables, which state which sequence of stops make up a route pattern, and route tables, which state which patterns make up a route. The largest piece of data by far is the vehicle sensor data. All buses and trains record their speed, odometer value, and GPS position once every few seconds. Additionally, they record special events when they start or stop recording data, are assigned to a route, enter or exit what is called a "stop zone" around a station, come to a halt, or release or lock their doors. In total, there are about 350 active vehicles, 1,000 stations, and 90 routes. The recording spans one month and amounts to about 380 million lines of data.

Upon first examining the vehicle data, stored in just over 19,000 compressed text files, we noticed several inconsistencies. A considerable number of files appeared to be truncated and the checksum test of the compression format failed. For most of these files, however, the extracted content still ended with a "data collection complete" event, so no relevant data was missing. In addition to invalid lines of data caused by an incomplete compressed file, some correctly extracted lines specified invalid dates some 20 years earlier or invalid vehicle numbers, most likely caused by an invalid configuration or missing initialization of the vehicle. In the end, 89% of the data files passed our automated consistency checks.

Our first step was to filter the raw data, perform some basic consistency checks on the data lines, transform them into a binary format to speed up further processing, and store the result in several files each containing all data items of a certain type recorded within a given period of time and grouped by the vehicle that recorded them. To start out with a reasonable subset, we only extracted 24 hours worth of data and removed all speed and odometer items that did not change the current state of the vehicle (while vehicles were travelling at a constant speed or standing still). This amounted to about 300 MB of binary vehicle data. While having all data in memory is convenient at this stage of the development, we consider switching to a database system in the medium term to improve data scalability and externalize some of the processing.

2.1. Time and space

Our data items are ordered by vehicle number and time stamp and apart from data inconsistencies and resetting clocks, a vehicle can have only one state and be at only one location at any point in time. We can easily draw a diagram that shows the speed of a vehicle over time and we can overlay this diagram with information on what route a vehicle is currently serving and whether it is within the stop zone of a station. We can determine the current vehicle state for a given point in time by looking up the corresponding item index and searching backwards for an item of the appropriate type. To accelerate this process, we keep speed and odometer values, positions, and other events all in separate lists, so we do not have to search through a lot of speed changes to determine what route a vehicle is serving at a given point in time.

In the same way, we can retrieve the GPS position of a vehicle, project that onto a two-dimensional map display and paint a track or animate a position indicator to visualize the vehicle movement in space and time. Vehicles in public transport, however, usually do not move freely around a map but follow specific, predetermined routes. This is especially true for trains and trams, which are bound to follow their rail tracks, and it also holds for buses, which follow the roads from stop to stop. Therefore, one important alternative view of space is that of a strictly linear progression along a predetermined path. Diagrams that map speed or some other value to the travelled distance are much more meaningful than they would be for an arbitrary movement, as for different vehicles travelling along the same path, the same distance from some set point of origin refers to the same point in space. Overlaying the data curves of different vehicles by travelled distance rather than time can reveal peculiarities that are specific to a certain location.

In order to create such a display, we need to be able to access our data items by travelled distance rather than time. The movement of a vehicle can be seen as a series of segments travelling from one station to the next. The departure station is the point of origin and a vehicle should reach the destination station after travelling the distance between them. When a vehicle stops at a station, we store its odometer value. Once we know the odometer values at both ends of a segment, we can map any position between the respective stations to a odometer value for the vehicle. Searching for a data item with this odometer value, we can find the corresponding time and use that to find any other related data items.

While a vehicle will only have one set of data items for a given point in time, it will usually reach the same point in space multiple times. Data sequences indexed by location therefore have multiple "runs", i.e. multiple values per location belonging to multiple passes by the same vehicle. To account for the fact that there may be multiple paths between two stations, we compare the distance actually travelled by the vehicle to the nominal station distance according to the infrastructure data and discard vehicles that seem to have deviated from the expected path.

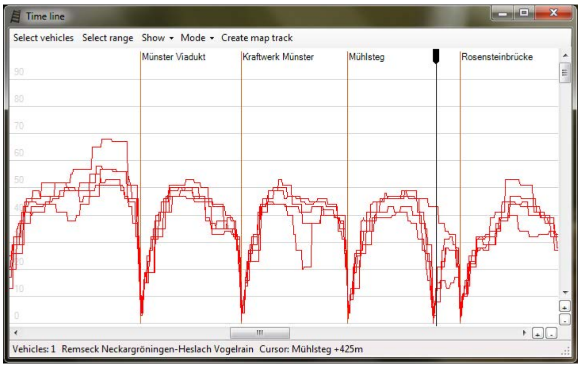
3. Visualization

We currently use three principal methods of displaying the data: A table view, a diagram view, and a map view. The views can be opened and operated separately, but selections in one view are reflected in other views, so we adhere to the common multiple views and linking and brushing paradigms.

3.1. Table view

In most visual analytics applications, it is desirable to have a separate view that displays the actual data as it is, without any aggregation or complex visualization. More sophisticated views enable users to discover points of interest and come up with hypotheses, but they might want to turn to a raw data view to verify their findings or examine them in detail. We supply a data table view that displays the data items of one or multiple vehicles along with their time stamp, type, and any additional parameters. To facilitate using the view, we reflect the type of the data items in the background colour of the data rows and offer an option to filter the displayed items.

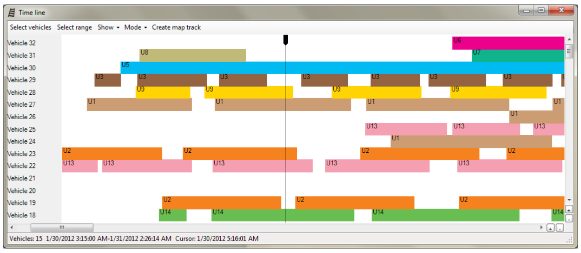
3.2. Diagram view



Vehicles: 1 Remseck Neckargröningen-Heslach Vogelrain Cursor: Mühlsteg +425m

Figure 1: *The diagram view displaying the vehicle speed during multiple trips of a single vehicle along the stations of a route (route pattern range).*

The diagram view (Figure 1) shows a graphical representation of one or more data sequences across a data range. We currently support two types of data ranges: Time ranges and route pattern ranges. With a time range, the diagram shows the development of one or more data sequences over time. Multiple vehicles and multiple data sequence types (such as current speed and halt events) can be displayed simultaneously, but there will never be more than one value for any combination of vehicle, data sequence type, and time. Multiple sequences can either be overlaid, which is useful, for example, for comparing the speed graphs of multiple vehicles, or stacked, which is useful, for example, for getting an overview over which vehicles serve which routes over the course of a day (Figure 2).



Vehicles: 15 1/30/2012 3:15:00 AM-1/31/2012 2:26:14 AM Cursor: 1/30/2012 5:16:01 AM

Figure 2: *Stacked bars showing which vehicles served which routes (time range).*

A route pattern range (as used in Figure 1) is defined by a sequence of pattern segments, each specifying a departure and an arrival station. We map the items of a data sequence to this linearized spatial axis with the method we described in Section 2.1. Unlike a time range, a route pattern range can contain multiple data values for a given vehicle and position. Plotting the speed of a single vehicle will usually result in a number of overlapping graphs, which can be used to compare multiple trips along the same route.

Once users have selected a set of vehicles and data sequences and specified a data range, they can zoom and pan the view. Clicking into the diagram will place a vertical line representing the cursor position. If a time range is active, the diagram view shows the cursor time in its status bar, the map view shows the locations of all selected vehicles at this point in time, and the data view scrolls to the corresponding item in its list of data items. For a range view, the cursor position is given as the distance to the previous station and vehicles are plotted at the corresponding location on the map view.

To aid in analysing over-plotted graphs, the user can add an average graph. If the display is configured to display the speeds of vehicles between two stations, for example, this average graph shows the average speed of vehicles at a certain position and the user can easily see whether a given vehicle is above or below the average speed at this location. Additionally, the user can select to display the standard deviation as well, overlaying the vehicle graphs with both an average value and a range of common values. In their combination, these allow for some interesting observations: Local conditions that affect all vehicles, such as tight corners or general speed limits, can be seen in the graph of the average speeds (Figure 3 top). Conditions that affect a considerable number of vehicles, such as intersections or traffic lights which may or may not force a vehicle to slow down or stop, cause visible bulges in the standard deviation tube (Figure 3 middle). If single vehicles deviate from the common path, these can be identified as outliers which create singular graphs well outside the standard deviation marks (Figure 3 bottom). To further investigate these outliers, the user can click on a graph to open the data view and scroll to the corresponding vehicle and position or time.

To reduce the visual clutter caused by overlaying a large number of graphs, most of which roughly follow the average graph, we provide an option to fade out graphs while they are near the average and only draw them fully opaque when they keep a certain distance, measured in multiples of the standard deviation. This ensures that both the average progression with its standard deviation and any outliers are clearly visible.

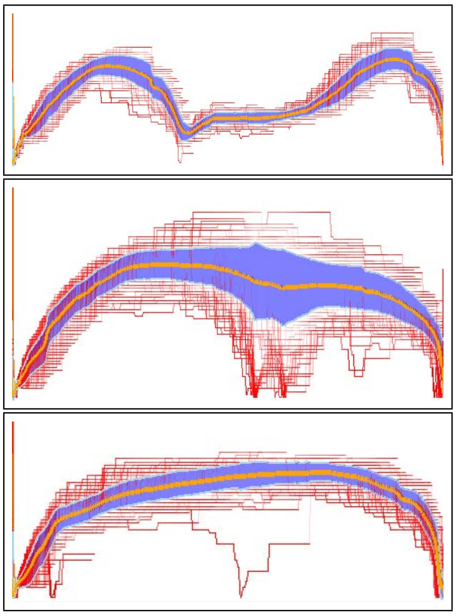


Figure 3: *Top to bottom: A characteristic average speed (orange), a noticeably increased standard deviation (blue) where some vehicles stop and some do not, a single outlier of a vehicle stopping while all others do not.*

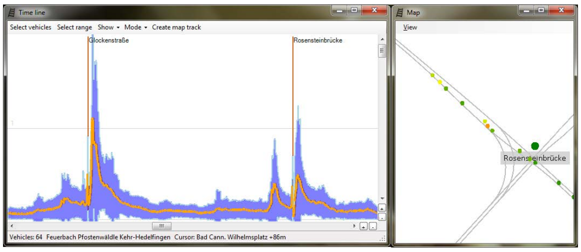
3.3. Map view

The map view is a two-dimensional display of stations and rail tracks (Figure 4 right). We received the rail infrastructure as a set of tracks. There is no connectivity information between tracks and no reference to the logical routes or stations other than a short textual description. The user can use the map to display data in a geographical context. When brushing over a graph in the diagram view, the corresponding geographical position of the vehicle or vehicles is highlighted on the map and one can follow the progress of the vehicle across the city. A diagram graph can be turned into a map track, displaying the values of the graph as colour-coded blobs at their geographical location. Average and standard deviation graphs can be turned into map tracks as well, allowing for a quick look at the geographic context in which certain values occur.

4. Derived data sequences

When mapping data items to a visual display, we are not limited to simply using the value of one of the attributes but can calculate derived values that are not present in the original data. For example, the discrete differentiation of the speed sequence yields the acceleration of the vehicle. This may be significant for assessing passenger comfort. The second derivative of the speed is the jerk, i.e. the rate of change of the acceleration. Sudden changes in acceleration reduce passenger comfort, so an analyst may be interested in whether there are certain locations within the transport network that are prone to causing jerks.

Jerks are to be expected when arriving at or departing from a station. Displaying the average jerk along a route can reveal locations where jerks occur outside a station. Figure 4 shows one such situation. The map view places these readings in a geographical context. Checking the location on a city map shows that there is both a street and a pedestrian crossing just before the station, so these jerks may be caused by conflicts with other traffic.



Vehicles: 64 Feuerbach Pfostenwäldle Kehr-Hedelfingen Cursor: Bad Cann. Wilhelmsplatz +86m

Figure 4: *A display of the average jerk (orange curve) and its standard deviation (blue) helps in discovering a location where vehicles experience a distinct average jerk while not near a station. The blobs on the map show both the location and the colour-coded intensity of jerks.*

5. Outlook

We have only just begun to explore the data set. Future versions of our tool will explore additional ways of visualizing the data and customizing the displays. A purely visual analysis is unlikely to find larger scale patterns, so we intend to include more analysis tools that will aid the user in finding facts of interests. Clustering algorithms could be used to identify different classes of trips, and more aggregation methods such as median values could improve the line plots. An automatic outlier detection could point the user at trips that differ from what is common for a certain route. Extending the analysis to longer time periods, we will have to implement out-of-core techniques or use a database system to ensure data scalability.

We intend to present our work to our contacts at the transportation company and get some expert feedback on which of these methods provide a novel benefit to the domain and which results could have just as easily been found using the tools they already have in order to evaluate the usefulness of visual analytics in the public transport domain.

Acknowledgements

The data set and the data specificiation were kindly provided by Stuttgarter Straßenbahnen AG and Trapeze Switzerland GmbH.

References

[AA05] Andrienko G., Andrienko N.: Visual exploration of the spatial distribution of temporal behaviors. *Information Visualisation, 2005. Proceedings. Ninth International Conference on* (2005), 799–806. 1

[AA08] Andrienko G., Andrienko N.: A visual analytics approach to exploration of large amounts of movement data. *Visual Information Systems. Web-Based Visual Information Search and Management 5188* (2008), 1–4. 1

[BBBL11] Boyandin I., Bertini E., Bak P., Lalanne D.: Flowstrates: An approach for visual exploration of temporal origin-destination data. *Computer Graphics Forum 30*, 3 (2011), 971–980. 1

[KVJ08] Kane L., Verma B., Jain S.: Vehicle tracking in public transport domain and associated spatio-temporal query processing. *Computer Communications 31*, 12 (2008), 2862–2869. 1

[LJF08] Lundblad P., Jern M., Forsell C.: Voyage analysis applied to geovisual analytics. *Information Visualisation, 2008. IV '08. 12th International Conference* (july 2008), 381–388. 1

[SWvdWvW11] Scheepens R., Willems N., van de Wetering H., van Wijk J. J.: Interactive visualization of multivariate trajectory data with density maps. *Pacific Visualization Symposium (PacificVis), 2011 IEEE* (2011), 147–154. 1

[TK09] Thomas J., Kielman J.: Challenges for visual analytics. *Information Visualization 8*, 4 (2009). 1

[vFHHP09] von Ferber C., Holovatch T., Holovatch Y., Palchykov V.: Public transport networks: empirical analysis and modeling. *The European Physical Journal B - Condensed Matter and Complex Systems 68* (2009), 261–275. 1