Visual Fusion of Mega-City Big Data: An Application to Traffic and Tweets Data Analysis of Metro Passengers

Masahiko Itoh*, Daisaku Yokoyama*, Masashi Toyoda*, Yoshimitsu Tomita†,
Satoshi Kawamura‡ and Masaru Kitsuregawa§
*the University of Tokyo, Email: {imash,yokoyama,toyoda}@tkl.iis.u-tokyo.ac.jp
†Tokyo Metro Co., Ltd, Email: y.tomita@tokyometro.jp
‡Tokyo Metro Co., Ltd., the University of Tokyo, Email: s.kawamura@tokyometro.jp
§National Institute of Informatics, the University of Tokyo, Email: kitsure@tkl.iis.u-tokyo.ac.jp

***Abstract*—Transportation systems in mega-cities are often affected by various kinds of events such as natural disasters, accidents, and public gatherings. Highly dense and complicated networks in the transportation systems propagate confusion in the network because they offer various possible transfer routes to passengers. Visualization is one of the most important techniques for examining such cascades of unusual situations in the huge networks. This paper proposes visual integration of traffic analysis and social media analysis using two forms of big data: smart card data on the Tokyo Metro and social media data on Twitter. Our system provides multiple coordinated views to visually, intuitively, and simultaneously explore changes in passengers' behavior and abnormal situations extracted from smart card data and situational explanations from real voices of passengers such as complaints about services extracted from social media data. We demonstrate the possibilities and usefulness of our novel visualization environment using a series of real data case studies about various kinds of events.**

I. Introduction

Public transportation systems, such as railways and metros, in mega-cities are always required to increase their resilience to extreme situations caused by various events. For instance, Tokyo, which is the biggest mega-city in Japan, will host the 2020 Summer Olympics, which will cause large scale movements of people over the wide area around Tokyo. Powerful inland earthquakes are also estimated to possibly occur in the Tokyo metropolitan area. Public transportation systems are now preparing responses for these events.

To increase the resilience of the systems, lessons must be learned from past events to understand how the systems are affected by changes in passengers' behaviors. Integration of smart card data and social media data enables us to replay past events and to discover abnormal situations of transportation systems, propagations of abnormalities over transportation networks, and passengers' complaints or dissatisfaction about which even train system operators and station staff do not know.

Our final goal is to implement the system for real-time analysis and prediction of passenger behaviors in a complex transportation system using real-time transportation logs and social media streams. It can support the evidence-based improvements of customer services such as optimal operation of transportation systems, navigation of passengers, and allocation of sufficient staff, trains, and other resources. It would also help offer proper staff training. Moreover, it can adapt to digital signage advertising to select appropriate advertisements for passengers going to gatherings such as sporting events or concerts.

As the first step to this goal, this paper proposes a novel visual fusion analysis environment that can support ex post evaluations of trouble in a metro system by using two forms of big data: archived transportation logs from the smart card system of the Tokyo Metro and social media data from Twitter. Knowledge acquired through the visualized results mostly reflects real situations such as disasters, accidents, and public gatherings.

For supporting effective exploration, the environment needs to satisfy the following requirements:

1) Discovering unusual phenomena from the wide range of temporal overviews that are derived from differences between daily and event-driven passenger behaviors. The techniques for intuitively verifying effects of known events and discovering trouble unknown to even train system operators are desired.
2) Understanding changes in passenger flows and spatial propagation of unusual phenomena in each time period on a wide area metro network. A visual exploration environment is necessary to intuitively understand the route, speed, and range of propagation of the unusual phenomena such as abnormal crowdedness. These are difficult for the train system operators to understand because the transportation system network in Tokyo is extremely dense and complicated.
3) Exploring reasons for unusual phenomena or their effects from real users' voices. A system is required for exploring information about passengers' complaints, activities such as use of taxis or buses, and confusing situations in stations, which often cannot be obtained from customer support or operation trouble databases.

To meet these requirements, we built a novel visual exploration method integrating the following visualization techniques:

978-1-4799-5666-1/14/$31.00 ©2014 IEEE

1) HeatMap view provides a temporal overview of unusual phenomena in passenger flows. It is used for spotting interesting phenomena by using patterns of colors. Although it does not provide spatial context, after finding out interesting temporal spots showing crowdedness or emptiness, we can explore spatial changes in them by combining HeatMap views with AnimatedRibbon views. Moreover, we can observe their causes and effects by combining HeatMap views with TweetBubble views.

2) AnimatedRibbon view visualizes temporal changes in passenger flows with spatial contexts and propagation of unusual phenomena over the whole metro network using animation. AnimatedRibbon view dynamically visualizes temporal changes in passenger flows by using animated 3D stacked colored ribbons for both directions of flows. It can simultaneously visualize two attribute values for the directed flows such as absolute counts (the scale of passenger flows) and relative counts (deviation from average) of passengers.

3) TweetBubble view provides an overview of trends of keywords explaining the situation during the unusual phenomena. It enables us to explore what passengers saw, heard, and felt in the situation by selecting stations or lines and dates through HeatMap view or AnimatedRibbon view. It uses a bubble chart to represent the popularity of important words and Sparklines [1] to show time-series of the appearance of each keyword in each hour for finding bursting timing.

We demonstrate the usefulness of our novel visualization system through a series of case studies extracted from real data related to natural disasters, accidents, and public gatherings. These case studies show how our visualization system enables users such as domain experts in the metro operating company and metro passengers to explore hidden knowledge based on data-driven analysis and visualization that were previously unattainable.

In what follows, we give an outline of related work in Section II. We offer information about our data set in Section III. We then describe a method for extracting passenger flows in Section IV and situational explanations in Section V. We introduce our novel exploration environments in Section VI. We present some case studies in Section VII. This article ends in Section VIII with a conclusion.

II. Related Work

A. Smart Card Data Analysis

Smart card data is one of the data sources to analyze operation of public transportation systems [2], [3]. Ceapa et al. focused on congestion patterns of some underground stations in London to reveal station crowding patterns to avoid traffic crowdedness [4]. They utilized data of oyster cards, the smartcards used on the London Underground. Their spatio-temporal analysis showed a highly regular crowding pattern during the weekdays with large spikes occurring in short time intervals. Sun et al. provided a model to predict spatio-temporal density of passengers and analyzed it for one MRT line in Singapore [5]. However, previous work only focused on a single selected line or some stations. One reason is that most smart card data does not include transfer station information. Our work speculates the most probable path of each trip from origin and destination in smart card data and succeeds in visualizing propagation of effects of trouble on the metro network. As far as we know, there has been no research on the visualization of such propagation of influences spreading over a wide range of public transportation systems such as metro networks.

B. Spatio-temporal Information Visualization

There have been some research on and systems developed for the visualization of geo-spatial and temporal values on a map. Tominski et al. introduced 3D icons into a map for representing spatio-temporal data [6]. Each 3D icon, which emphasizes linear or cyclic temporal dependencies, represents multiple time dependent attributes on maps. Thakur and Hanson also used 3D icons on maps to represent spatio-temporal distributions of time-varying quantities in a single view [7] and provided 2D icons called Data Vases to represent the profiles of the time-dependent variable, in which the colors and the horizontal arrangement represent regional classifications. Their approach focused on describing changes in values at independent points and did not provide a method for representing temporal changes in values between two points or flows.

There has been some research on analyzing mobility data and extracting and visualizing important events or mobility patterns. Ferreira et al. provided a query model and a visual environment for exploring human activity patterns in New York City using taxi trip data [8]. Wang et al. extracted and visualized traffic jams and their propagation from data of taxi trips in Beijing [9]. Andrienko et al. extracted and characterized important places from mobility data such as GPS tracks of cars and flight trajectories and visualized them in 3D spatio-temporal space [10], [11]. Unlike these cases, smart card data in our system includes origin-destination (OD) data without trajectory information. We therefore need to speculate the most probable route for each trip from OD data and visualize aggregated passenger behaviors.

Tominski et al. showed the usefulness of 3D trajectory bands to visualize trajectory attribute data [12], [11]. In their visualization, attribute data of individual trajectories was visualized as color-coded bands and sets of trajectories were visualized by stacking the bands. Cheng et al. also utilized 3D staked bands to represent overview of spatio-temporal changes in attribute data on a road network [13]. Stacked and color-coded 3D bands are useful for representing spatio-temporal changes in an attribute value on the map, but they cannot represent two or more kinds of attribute values or their scale such as the number of people. Our approach utilizes 2D heatmaps for overviewing temporal changes in flows and 3D animated ribbons for simultaneously visualizing changes in absolute counts such as the number of passengers and relative counts such as the deviation from the average and how these propagate in a complicated network.

C. Spatial Tweet Visualization

LeadLine [14] detected events from social media data, extracted information about 4 Ws (who, what, when, and where) related to the events, and then visualized the information in coordinated views. SensePlace2 [15] provided an integrated environment for filtering and visualizing space-time-theme information from twitter streams. Their approaches focus on exploring events from social media data without using other data resources.

Pan et al. provided a system for traffic anomaly detection from human mobility data and anomaly analysis using social media data [16]. They used term clouds to visualize terms related to the detected anomalies. Although in their approach visualization is only used for showing detected results, our work focus on providing interactive environments for finding anomalies and exploring them in detail by using two forms of data from smart card system and social media.

III. Datasets

A. Smart Card Data

We use a large scale data set of travel records from March 2011 to May 2014 on the Tokyo Metro extracted from the smart card system. Tokyo has the complicated train route map[1]. It consists of lines of various kinds of railway companies including Tokyo Metro, Toei Subway, Japan Railway (JR), and many private railroads. We analyze large scale log data covering almost all of the business area of Tokyo. It consists of 28 lines, 540 stations, and about 350 million trips. This includes lines and stations besides Tokyo Metro ones if passengers used lines of other railway companies for transfers.

In our experiments, we use passengers log data from anonymous smartcards without personal identity information, such as name, address, age, and gender. Card ID is eliminated from each record. Each record consisted of the origin, destination, and exit time[2]. Since transfer information was not included, we estimated the probable route for each trip (as explained in Section IV).

Trains in Tokyo are mainly used by working people, so the usage patterns of trains on weekdays and weekends may be different. We separate the data into weekdays and weekends and analyze them independently. National holidays and some other days in vacation seasons are treated as weekends.

Passengers are expected to behave with some periodic patterns, especially daily ones, thus we try to do a statistical analysis of this data. Figure 1 shows the average and standard deviation of the number of passengers at every time period of the day through one year, from Apr. 2012 to Mar. 2013. The error bar of each point indicates standard deviation. To extract Figure 1, we first estimate the trip time length of each trip log (mentioned in Section IV) and then accumulate the number of passengers who were travelling at a certain time period. The time periods are divided every 10 minutes. Weekdays and weekends have clearly different demand patterns. The deviations of weekdays are considerably smaller than those of weekends. This means that most passengers actually behave in a periodic manner, so we may be able to detect some irregular accidents or events by comparing the differences with the average number of passengers at each section. We try to confirm this hypothesis in the following sections.

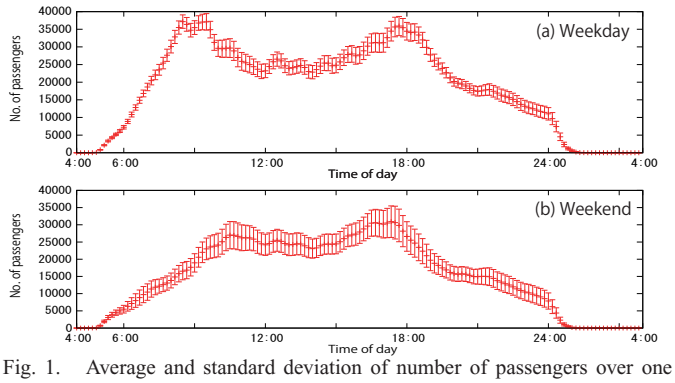


Fig. 1. Average and standard deviation of number of passengers over one year (Apr. 2012 to Mar. 2013): (a) weekdays, and (b) weekends and national holidays.

B. Social Media Data

Social media immediately reflects real world events such as accidents. In this paper, we utilize Twitter as a social media data resource. We have been crawling through more than three years' worth of Twitter data from Twitter API from March 11, 2011. We started crawling through data from famous Japanese users. We first obtained timelines of these users and then repeatedly expanded the set of users by tracing retweets and mentions in their timelines. We then obtained data of more than 2 million active users and 18 billion tweets.

IV. Extraction of Passenger Flows

With the recorded smart card data, we can understand how many passengers used a certain station. However, that data does not include the entrance time, so we could not see how many passengers are there within a certain time period. Moreover, if we try to estimate the crowdedness of each train, or effects of an accident at a certain location, the origin-destination pair is insufficient. We must figure out the travel path of each passenger for such requests.

A. Estimating Daily Passenger Flows

There are several possible paths to take from an origin station to a destination station. A smart card log contains information about where a passenger touched in and where and when he/she touched out. It does not include the entrance time or transfer stations' information. We therefore speculate the most probable path for each trip (origin and destination pair) by assuming that they take the shortest time path.

We assume that total travel time (t) of each trip is defined as $t = T + C + W$, where:

- T is the time while passengers are riding trains. It defined by using the timetable.
- C is the walking time while passengers are transferring trains. It relates to the structure of the station, so it differs

[1] http://www.tokyometro.jp/en/subwaymap/index.html
[2] No records contain trip start times.

at every station. We roughly define these times by using the information from the train company.

- W is the time waiting for a transfer. We define this as (average train interval / 2) extracted from the timetable. It differs on every line.

With this model, we can calculate the estimated travel time of any travel path. We then search for the shortest time path of every origin-destination stations pair by using the Dijkstra algorithm.

We want to find unusual phenomena that differ from the usual cyclical patterns of the passengers. For this purpose, we first estimate in which section of a line a passenger passes in a particular time period from the speculated shortest time path and exit time. We then accumulate the number of passengers who travelled a certain section in a certain time period (every 10 minutes or one hour). Data for weekdays and data for weekends are separately analyzed because weekdays and weekends show clearly different patterns as shown in Section III-A. After that, we calculate the simple moving average (SMA) of the previous one year for each month and calculate standard deviation using the same time window. SMA reflects daily cyclical patterns, and unusual patterns can be detected by comparing it with log data.

All passenger flows from one-day smart card records take several minutes to estimate using one server. Many parts of our current system are experimentally implemented and have large room for improvement in terms of the execution performance.

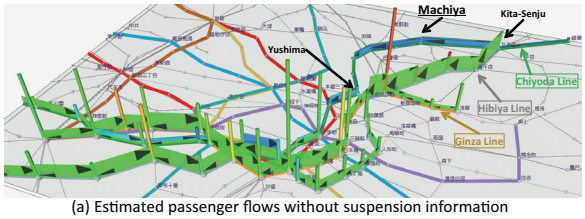
B. Estimating Passenger Flows after Accidents

Accidents sometimes cause service suspensions at several sections, making passengers take detours. In such situations, the route estimation method proposed in Section IV-A cannot calculate an appropriate route because the shortest path would be changed by service suspensions.

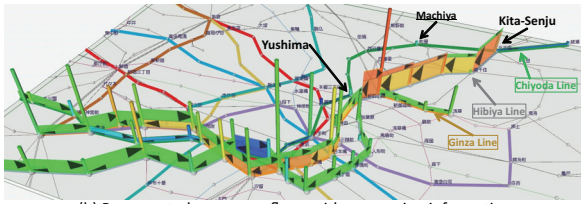
We can recompute the shortest paths considering the suspension information such as suspended sections and time. This refines passenger flow estimation to make it more appropriate to describe what happened at that time. We provide interfaces to input constraints of suspended lines and/or sections and start and end times of suspension on our visual exploration environment shown in Section VI. The visual exploration environment visualizes the recomputed result. We can then visually check how passengers take detours and concentrate on particular lines.

Figure 2 compares passenger flows with and without suspension information on 27 Nov. 2013, the day on which an accident resulting in injuries happened at Machiya station on the Chiyoda Line. Figure 2 (a) shows passenger flows without suspension information. Figure 2 (b) shows recomputed passenger flows using suspension information based on factual information. In Figure 2 (a), the flow of passengers continues to exist even after the Chiyoda Line service is suspended. In Figure 2 (b), sections from Kita-Senju to Yushima are suspended from 9:59 to 10:37. We can find out that passenger flows concentrate on the Hibiya and Ginza Lines to avoid selected sections on the Chiyoda Line[3].

We can use suspension information obtained from an external information resource such as the metro operating company or the transport information webpage as inputting constraints for our system. The metro operating company holds information about events that disrupt their subway system. We also collect the train operating condition information from the transport information webpage of a third-party company [4].



(a) Estimated passenger flows without suspension information



(b) Recomputed passenger flows with suspension information:
Sections from Kita-Senju to Yushima on the Chiyoda Line are suspended from 9:59 to 10:37.

Fig. 2. Passenger Flows after accident on 27 Nov. 2013, the day on which an accident resulting in injuries happened at Machiya station on the Chiyoda Line.

C. Discussion

Our passengers flow model is constructed on the assumption that every passenger will take the fixed shortest path. Their real behavior is more diverse; they will also consider travel fees, crowdedness, ease of train transfer, etc. Using a probabilistic behavior model may be more appropriate to reflect such diversity. When considering the cases of accidents, our method makes another train scheduling assumption: trains that were unaffected by the accidents keep travelling on time. Trains were stopped at certain sections in the case of serious accidents, but more modest actions such as delaying trains or partially eliminating services would have happened in many cases. We have already tried to construct a preliminary behavior model of passengers after accidents have happened [17] and plan to improve the model by introducing such details.

Since the current smart card system does not have the information of entrance time or transfer points of each trip, we cannot evaluate the preciseness of the estimated passenger flows directly. We interviewed some of the staff of the train operator and found that the extracted flow seems to correspond to their knowledge of the daily operation. We plan to evaluate the preciseness through comparison with other statistical survey results such as traffic censuses.

V. Extraction of Situational Explanation

Social media enables people to post information about what they saw, thought, and did during and after events such as

[3]The meaning of each visual element is described in Section VI.
[4]http://transit.goo.ne.jp/unkou/kantou.html (in Japanese)

accidents. We can extract more precise or fine-grained information about the events that sometimes cannot be obtained by operating companies.

We extract a set of words (weighted by word frequencies based on the measure similar with tf-idf) for overviewing and explaining situations. For this purpose, we first calculate word frequencies for every co-occurring word for each station name or line name on each date and time from the data set described in Section III-B as $tf(word, station/line, date\ and\ time)$ (if we specify a start time and an end time, we use the sum of the word frequency between them ($tf(word, station/line, timewindow)$)).

We then count the number of days when each word appears for each station or line and treat it as $df(word, station/line)$. In this case, we treat a set of tweets on one day including the name of a station or line as one document for each station or line. It is used for decreasing importance of words commonly used all the time for each station or line. Small accidents or short delays happen almost every day around Tokyo. Therefore, if we use df for all documents that are related to all stations and lines, words related to trouble may not be treated as important. Moreover, the characteristics of co-occurring words that appear routinely are different among stations or lines. Therefore, we calculate df for individual stations or lines.

We finally calculate $weight(word, station/line, date\ and\ time/timewindow)$ as $tf \times idf(word, station/line)$ (s.t. $idf = \log(\frac{|date|}{df(word, station/line)}) + 1$).

VI. Exploration Environment for Passenger Flows

We provide three types of visualization views: HeatMap view and AnimatedRibbon view to explore passenger flows and spatio-temporal propagation of crowdedness or emptiness extracted by the methods mentioned in Section IV, and TweetBubble view to explore situational explanations extracted by the method described in Section V.

HeatMap, AnimatedRibbon, and TweetBubble views are coordinated with each other. For example, users can select time stamps and lines on HeatMap view, and then AnimatedRibbon view starts animated changes in values for selected lines and time or TweetBubble view represents trends related to the selected lines and time.

A. HeatMap View

HeatMap view shows an overview of temporal crowdedness or emptiness to help users to easily discover unusual phenomena and explore their temporal characteristics. For these purposes, HeatMap view provides functions for showing deviation from average passenger flows in each time bin on each section for every line over one day or one month.

We provide two types of HeatMap view: one for the monthly overview and the other for the daily overview. It uses the x-axis for the timeline and the y-axis for lines. The timeline is divided every 1 hour in monthly mode (Figure 3) and every 10 minutes in daily mode (Figure 4).

It enables users to find unusual phenomena on a particular line, over multiple lines, or for several days such as in Figure 3. Figure 3 shows a HeatMap view for March 2013 in monthly mode. The Tokyo Metro Fukutoshin Line started to directly connect to another private railroad at Shibuya station on 16 March 2013. We can find that it caused dramatic changes in passengers' behavior. The number of passengers on the Fukutoshin Line increased greatly after 16 March 2013.

Each line is represented by different colors, and both directions (up and down) are treated separately (Figure 4). Up and down lines are separated by color, and labeled by the starting stations as shown in Figure 4. Each up/down line consists of sections that are pairs of origin and destination stations as shown in Figure 4[5]. The order of lines can be manually changed. HeatMap view also can be zoomed and panned interactively. Users can interactively select lines or times to visualize in other views such as AnimatedRibbon view and TweetBubble view.

Although HeatMap view is useful for finding temporal characteristics of abnormal situations, it is still difficult for us to explore spatial characteristics and their propagation.

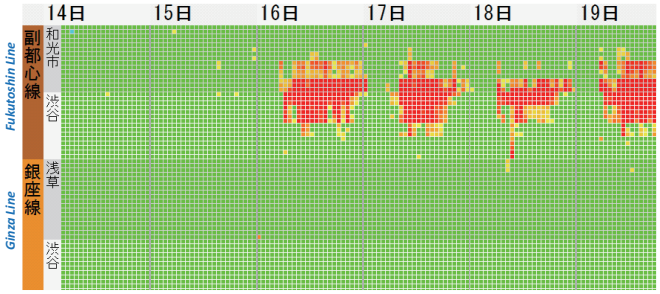


Fig. 3. HeatMap view for March 2013 in monthly mode in which the timeline is divided every 1 hour. On March 2013, the Fukutoshin Line started to directly connect to another line, causing a rapid increase in the number of passengers.

1) Color Encoding on HeatMap View: The color code for each cell in the HeatMap represents relative crowdedness or emptiness of each section compared with the average situation. For this purpose, we calculate z-scores (difference normalized by standard deviation) of each section for each time bin by SMA and standard deviations shown in Section IV. Red represents a higher z-score indicating crowdedness, blue represents a lower z-score indicating emptiness, and green represent a middle z-score that indicating a mostly normal situation.

Two types of thresholds (one for smaller value (S-th) and the other for larger value (L-th)) can be manually defined to emphasize small differences or change the range for viewing z-scores. For instances, Figure 5 (a) uses $S\text{-}th = 2.5$ and $L\text{-}th = 9.0$, and Figure 5 (b) uses $S\text{-}th = 2.0$ and $L\text{-}th = 5.0$. Z-scores for each block are normalized by using S-th and L-th, and then the color code is defined. If the absolute value of a z-score is smaller than S-th, then green is used. If the absolute value of a z-score is larger than L-th, the cell becomes

[5]We omit labels for origin and destination stations in other figures.

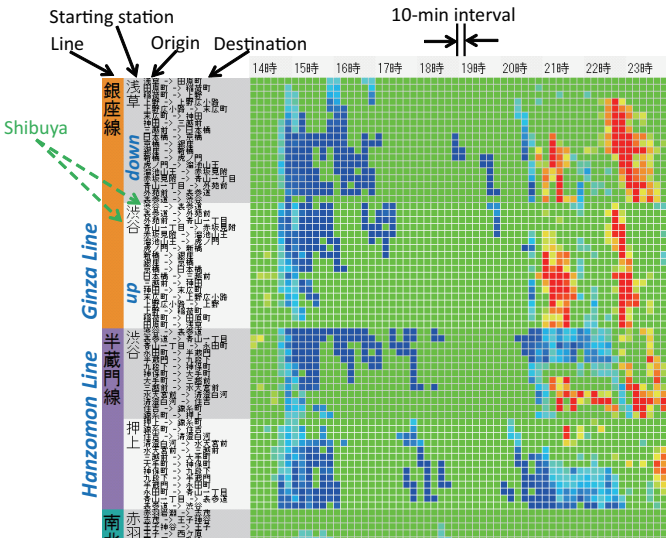


Fig. 4. HeatMap view 11 Mar. 2011, the day on which the Great East Japan Earthquake occurred, in daily mode. All lines were suspended just after the earthquake at 14:46 (shown as blue time bins). Some lines resumed around 20:40, causing concentration of passengers (shown as red time bins).

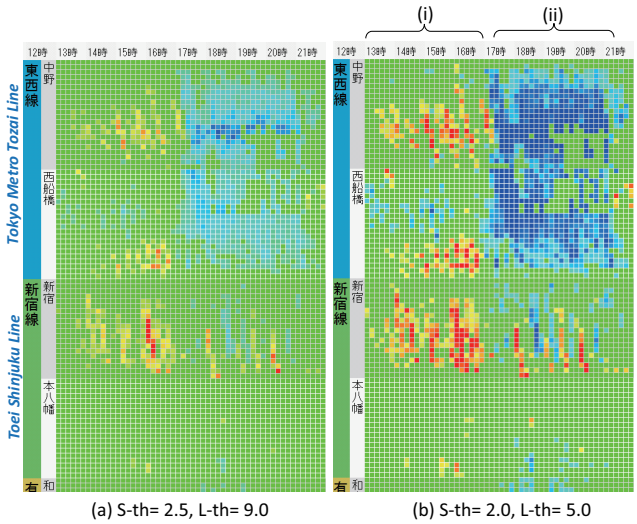


Fig. 5. HeatMap views related to the spring storm in April 2012 in daily mode with different threshold values (*S-th* and *L-th*): (a) for emphasizing only time bins with large z-score, and (b) for emphasizing small differences.

red/blue. Color is adjusted between green and red or blue. *S-th* and *L-th* values for specifying color code can be used for specifying colors in AnimatedRibbon view.

B. AnimatedRibbon View

AnimatedRibbon view visualizes animated temporal changes in the number of passengers and crowdedness or emptiness of each section in the Metro network. It dynamically shows changes in two attribute values (absolute number of passengers by using height of 3D ribbons and deviation from average by using color-coding) while maintaining geographical context in the Metro network. It is necessary for exploring the scale of passenger flows, particularly those with huge spikes, and exploring propagation of abnormal situations on the Metro network at the same time. Although there have been some studies using a heatmap [9] or wall map representation [12], [13] to represent temporal changes in a single kind of attribute value, as far as we know, there has been no research to simultaneously represent both absolute and relative values in flows in a network.

Figure 6 shows an example of AnimatedRibbon view. The number of passengers is represented by the height of stacked 3D ribbons, which consist of ribbons for two directions, on each section every 10 minutes. A 3D bar on each station presents the number of passengers who exited the station every 10 minutes. AnimatedRibbon view uses animation to dynamically display temporal changes in height and color of each 3D ribbon and 3D bar (as shown in Figure 6 (a) - (c)).

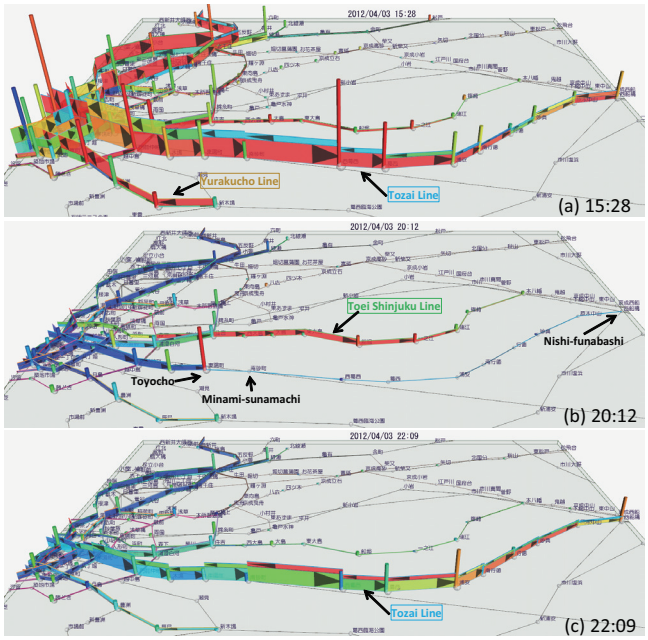


Fig. 6. Animated changes in passenger flows and propagation of crowdedness on AnimatedRibbon view related to the spring storm in April 2012.

1) Color Encoding for 3D Ribbons and Bars: The color of each 3D ribbon for each direction is defined using z-scores, *S-th*, and *L-th* specified in Section VI-A1. The color of each 3D bar also shows deviation, which is defined by z-score normalized by thresholds, from the average number of passengers who exited each station in the same way as passenger flows. In both cases, red represents higher than average, blue represents lower than average, and green represents the normal situation in the same way as HeatMap view.

We can also use transparency to represent z-score normalized by *S-th* and *L-th* defined in HeatMap view. In this case, ribbons that have a z-score lower than *S-th* can be hidden. This emphasizes important sections on which users should focus.

2) 3D Design Considerations: In AnimatedRibbon view, we simultaneously utilize heights in the 3D space for representing the number of passengers and colors for the level of crowdedness. Height is more suitable than color for repre-

senting values that have huge spikes such as the number of passengers shown in Figure 10 (b) because color does not have the dynamic range to permit extreme magnitude [18]. Utilizing 2D bands such as those used by Andrienkos [19] is one solution for representing the number of passengers. However, 2D bands would quickly suffer from severe overplotting [12] and the occlusion problem, especially around highly connected stations or caused by extremely big values. Dang et al. showed that utilizing heights in the 3D space to represent the magnitude of values in dense data area is useful to avoid overplotting instances [18].

Perspective foreshortening makes it difficult to compare the heights of ribbons and/or bars in the 3D space in different places from the camera. To avoid the problem, our system has an orthogonal projection mode in which the ribbons and/or bars in different places that are the same height look completely the same.

We utilize the metro network map on the basis of real geographical positions. The metro network illustrated in the AnimatedRibbon view is very complicated. Utilizing 3D ribbons and bars in such a complicated network sometime causes cluttering of visualization results and the occlusion problem. To avoid such occlusion, users can zoom, rotate, and pan the 3D space to interactively change the region being focused on. Moreover, users can hide ribbons or bars and select lines to show bars and ribbons on the selected lines. If sections consist of multiple lines, the number of passengers is the total number of them on the selected lines. Colors of bands can be made semi-transparent to see rear elements.

Users can also pan and zoom the route map in the 2D plane. There are some regions in the real route map where the station is extremely dense such as around Tokyo, Shinjuku, and Ueno stations. The occlusion problem of 3D elements easily occurs in such dense areas. Zooming the region in a 2D map is one solution to avoid occlusion in dense areas. However, we sometimes lose the overview of a wide area in zoomed route maps. We therefore implement a map distortion technique using fisheye view [20]. Users can see details in the dense area, which can be specified by interactively selecting a station or a point in the 2D map, and the overview of the surrounding area while maintaining geographical context to some extent (Figure 9 (a)). Users can specify arbitrary parameter values such as distortion factor d and range $Dmax$ defined by Sarker and Brown [20].

C. TweetBubble view

TweetBubble view shows an overview of aggregated words from people's tweets related to specified times and stations or lines, which can be selected by HeatMap view or AnimatedRibbon view, to easily explore the causes and effects of unusual phenomena.

In this view, the center node represents a selected station or line, and other nodes around the center node represent words co-occurring with the station or line name (Figure 7[6]). Each node holds tf value for each hour and df value described in Section V. We can interactively filter nodes (other than the center node) by total tf value of all hours in the day and time window using range sliders shown in Figure 7.

A TweetBubble view changes the size of nodes in accordance with the *weight* defined in Section V for the selected time window as $r\sqrt{weight}$ (r: constant). We adopt an automatic and dynamic graph layout algorithm based on a force-directed model [21] to visualize bubble charts. Nodes are colored differently in accordance with parts of speech (noun: green, verb: sky blue, adjective: pink).

TweetBubble view embeds Sparklines [1], which are small line charts, into every node to present variation of tf values for words over time (from 0:00 to 24:00). Parts of lines corresponding to the selected time window are highlighted in red.

We can read original tweets including the selected station or line name and word in the selected time window. The tweets are displayed in the bottom of the view by clicking an arbitrary node. These are sorted by time. Tweets are colored differently in accordance with their types (normal tweet: black, mention: blue, retweet: red).

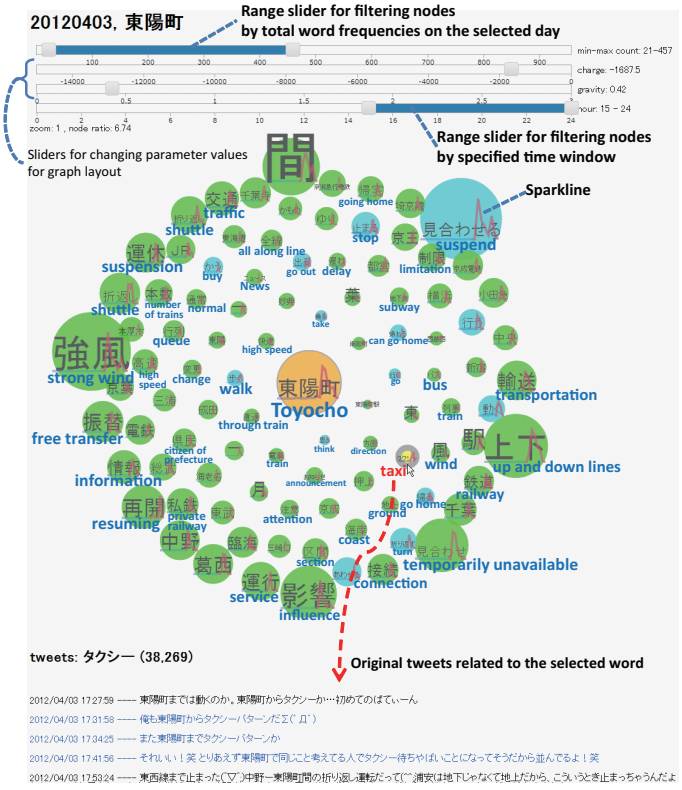


Fig. 7. TweetBubble view related to Toyocho station on 3 Apr. 2012, on which the spring strom came. It consists of a bubble chart and sparklines to represent importance of words and changes in appearance frequencies.

VII. Case Studies

In this section, we will demonstrate case studies using the extracted passenger flows in Section IV, the extracted situational explanations V, and the proposed visualization environments VI. These case studies show the usefulness of our

[6]We omit English labels for proper nouns or words that are too common.

system for exploring changes in behavior of passengers and influences of various kinds of events such as natural disasters (Section VII-A, and VII-B), accidents (Section VII-C), or public gatherings (Section VII-D).

We interviewed customer service staff of a train operating company and confirmed that the visualized results corresponded to their daily experiences. Moreover, the results of visualization may throw light on facts that even station staff did not know or give evidence for the situations that they understood somehow. Knowledge of phenomena obtained from the results is useful for optimizing usual operations and preparing for future disasters, accidents, and events.

A. The Great East Japan Earthquake

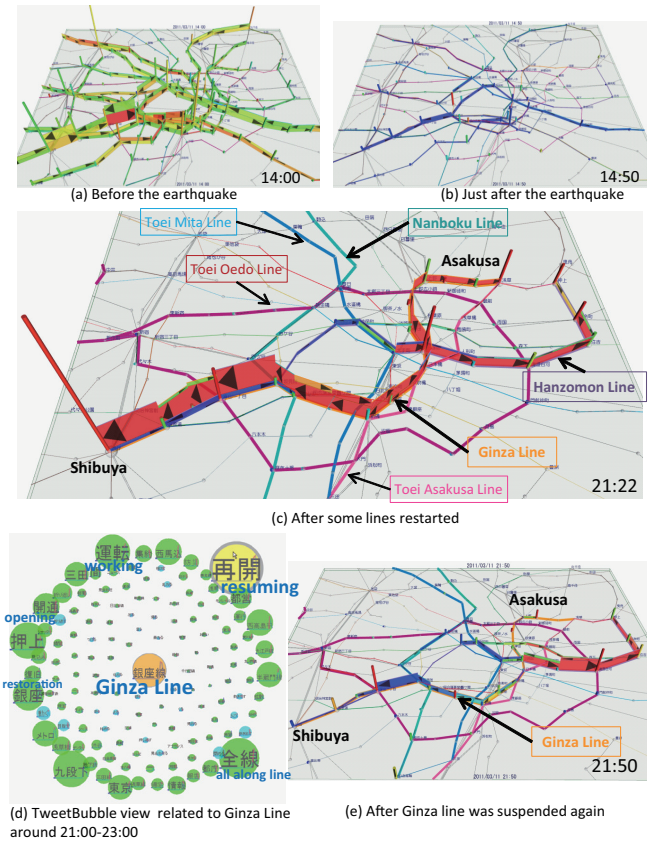
Figure 8 visualizes passenger flows on 11 Mar. 2011, the day on which the Great East Japan Earthquake occurred. The earthquake struck off the northeastern coast of Japan at 14:46. It had a seismic intensity[7] of 5-upper in Tokyo. A public report[8] notes that many public transportation systems suspended operation after the earthquake, so most people could not travel until midnight or the next morning.

Figure 8 (a) shows the situation just before the earthquake. We can find almost all lines were operating normally because their color is mostly green. We can see that almost all lines suspended operation after the earthquake from Figure 4 and Figure 8 (b). There are large blue areas in HeatMap view just after 14:46 in Figure 4. The color of each section turns blue in AnimatedRibbon view in Figure 8 (b).

The report stated that the Tokyo Metro Ginza Line and part of the Tokyo Metro Hanzomon Line resumed at 20:40. The Tokyo Metro Nanboku Line also resumed at 21:20. The Toei Oedo Line, a part of the Toei Asakuda Line, and a part of the Toei Mita Line resumed at 20:40, 21:20, and 21:15, respectively.

From the red areas shown in the upper-right part of Figure 4 and the red ribbons in Figure 8 (c), we can find a huge number of people were concentrated on the Ginza Line moving to Shibuya or Asakusa. We can explore the situation in which many people tweeted information such as "Ginza Line is running again" before and after it resumed as shown in Figure 8 (d). The spread of such tweets might have accelerated the concentration of people to the Ginza Line and Shibuya station.

We also found that the number of passengers who went to and exited Shibuya rapidly decreased around 21:50 after the concentration using AnimatedRibbon view in Figure 8 (e). Such rapid and short-term decreases cannot be shown in HeatMap in Figure 4. We searched for the reason by reading original tweets around 21:50 using TweetBubble view shown in Figure 8 (d). We then found many tweets such as "Ginza Line resumed once, but it is suspended again because of confusion at Shibuya station" from the tweets related to "resuming". The public report confirms this information was right.



(a) Before the earthquake (b) Just after the earthquake

(c) After some lines restarted

(d) TweetBubble view related to Ginza Line around 21:00-23:00 (e) After Ginza line was suspended again

Fig. 8. Visualizations of passenger flows and tweets on 11 Mar. 2011, the day on which the Great East Japan Earthquake occurred.

B. Spring Storm April 2012

The HeatMap view in Figure 5 and the AnimatedRibbon view in Figure 6 visualize changes in passenger flows on 3 April 2012, the day on which a spring storm that had the same intensity as a typhoon hit the Japanese mainland. Many companies in Tokyo urged employees to go home early that day. Figure 5 (b-i) and Figure 6 (a) show the Tozai Line, Toei Shinjuku Line, and Tokyo Metro Yurakucho Line became very crowded before the normal rush hours.

The Tozai Line suspended operation between Minami-sunamachi and Nishi-funabashi around 17:20. We can find many passengers exited Toyocho station in Figure 6 (b), and passengers started to use Toei Shinjuku Line to move to eastern areas in Figure 5 (b-ii) and Figure 6 (b). Red and blue stripes on the Toei Shinjuku Line in Figure 5 (b-ii) show it could not maintain normal operation. Many people therefore had no routes to take to eastern areas of Tokyo.

The Tozai Line resumed at 21:05. Figure 6 (c) shows passengers who had been left in central Tokyo started to move again on the Tozai Line after it resumed.

TweetBubble view in Figure 7 shows words related to Toyocho station from 15:00 to 24:00. Words shown as huge nodes mainly represent abnormal situations of service such as suspension and free transfer, or their causes such as strong

[7]http://en.wikipedia.org/wiki/Japan_Meteorological_Agency_seismic_intensity_scale

[8]http://www.mlit.go.jp/tetudo/tetudo_fr8_000009.html (in Japanese)

wind. These also include related words such as taxi, bus, and walk that represent passengers' real behavior, how they traveled from Toyocho to their destinations, during the storm. Original tweets including "taxi" are shown under the bubble chart. Most of these tweets said that there was a long line of people at the taxi stand. Such information is very important for understanding the influence of service suspension on activities of people to improve service operation, including cooperation with other transportation services.

Many typhoons pass through Japan every summer and autumn. Visualization results of situations during typhoons will be almost the same as the case shown in this section. People in the operating company had not been aware of such extremely confusing situations, especially in Toyocho station. These results gave them one piece of new evidence to help them discuss improving the transportation system around the east side of Tokyo.

C. Fire at Yurakucho on 3 Jan. 2014

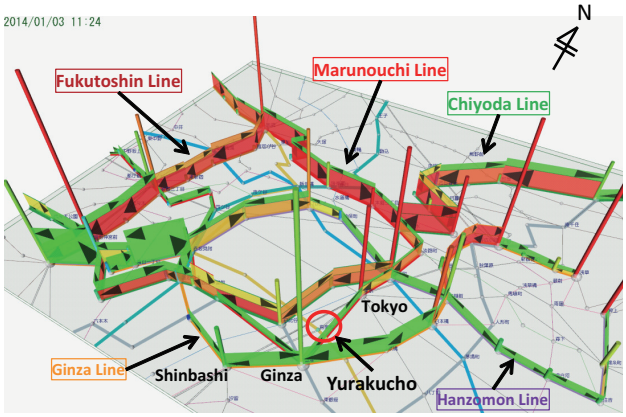
Figure 9 visualizes passenger flows after the fire around JR Yurakucho station on 3 January 2014. The fire started at around 6:30 a.m. and sent plumes of black smoke over Yurakucho station, which is an important gateway to famous business, shopping, and nightlife districts such as Yurakucho, Tokyo, Ginza, and Shinbashi (as shown in Figure 9 (a), which uses the distortion technique mentioned in Section VI-B2). It caused suspension of the JR Yamanote Line, JR Tokaido Main Line[9], and Keihin-Tohoku Line.

Figure 9 (b) and (c) show many people changed their routes to their destinations mainly by using the Fukutoshin, Marunouchi, and Chiyoda Lines. We can observe that many passengers switched to the Tokyo Metro Fukutoshin Line in place of the JR Yamanote Line in Figure 9 (b). The number of passengers increased mainly between Ikebukuro and Meiji-Jingumae. Many passengers switched from the JR Yamanote Line to the Chiyoda Line (from Kita-Senju to Meiji-Jingumae). Passengers changed to the Tokyo Metro Marunouchi Line to go to Tokyo station.

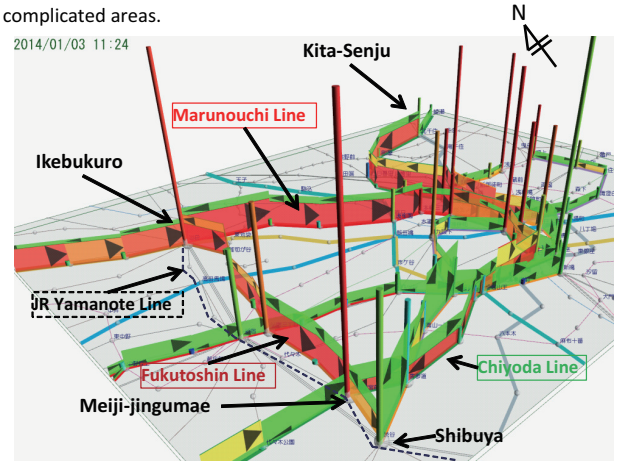
Figure 9 (c) shows changes in passenger flows on the Chiyoda Line from Kita-Senju station. Although many people normally transfer from the JR Joban Line to the JR Yamanote Line at Ueno station, they got off at Kita-Senju station and transferred to the Chiyoda Line to go to central Tokyo in this situation.

We have observed other examples of suspension of the JR Yamanote Line and confirmed similar changes in passenger flows in these situations. The examples shown demonstrate the influence of other railway companies' accidents on Tokyo Metro. Such indirect effects of accidents are hard to understand. This visualization results give us helpful evidence of and insight into the effects of such accidents on the basis of real data.

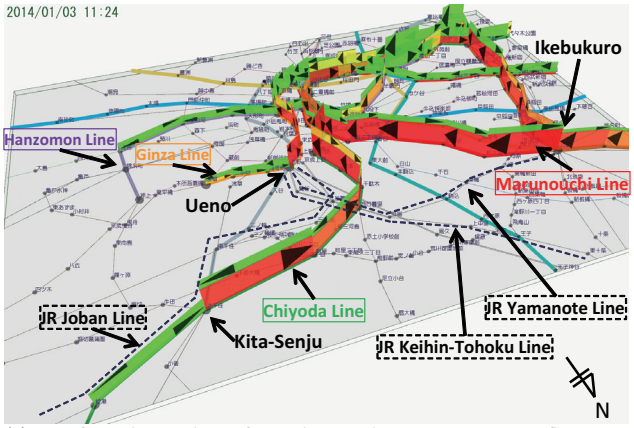
[9]The JR Tokaido Main Line runs from Tokyo, stops at Shinbashi, Shinagawa, and then eventually terminates at Kobe.



(a) View from the southeast for overviewing using distortion technique to show detail around JR Yurakucho station, which is one of the most complicated areas.



(b) View from the southwest for focusing on changes in passenger flows on Fukutoshin Line, Marunouchi Line, and Chiyoda Line.



(c) View from the northeast for explaining changes in passenger flows on Chiyoda Line.

Fig. 9. Effect on the fire around Yurakucho station on 3 Jan. 2014

D. A Parade by London Olympic Medalists in Ginza

Figure 10 visualizes changes in passenger flows on 20 August 2012, the day on which a parade by London Olympic medalists was held in Ginza. The parade lasted about 20 minutes from 11:00, and about 500,000 people gathered[10].

[10]http://www.joc.or.jp/english/londonolympics/parade.html

Figures show a massive amount of people gathered in Ginza before the parade started and quickly left from Ginza after the parade ended.

By using AnimatedRibbon view shown in Figure 10 (b) and (c), we can recognize extremely huge waves of passenger flows occurred before and after the parade. Figure 10 (a-i) and (b) show that many people moved toward the Ginza area from various quarters after 9:00. We can also find they started to leave Ginza just after the parade ended in Figure 10 (a-ii) and (c). This is a surprising result, because Ginza is one of the most famous shopping districts in Japan, but most people did not stay there for long. Figure 10 (c) shows that many passengers exited Shibuya, Shinjuku, Ikebukuro, Ueno, and Asakusa stations.

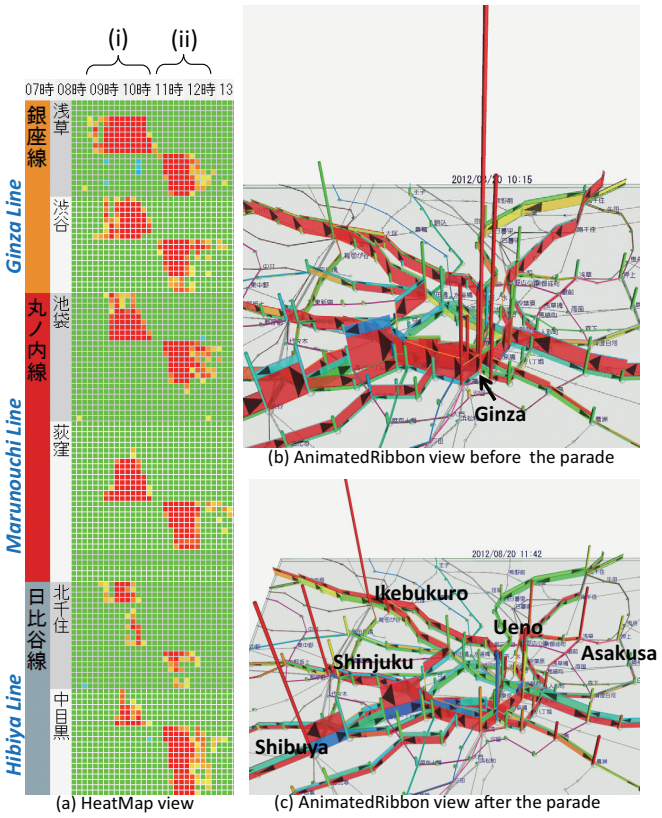


Fig. 10. Visualization of passenger flows on 20 August 2012, the day on which a parade by London Olympic medalists was held in Ginza.

VIII. Conclusion

We proposed a novel visual fusion environment to explore changes in flows of passengers on the Tokyo Metro and their causes and effects by using more than three years' worth of data extracted from the smart card system and Twitter. Our major contribution is a novel approach that extracts and visualizes (1) passenger flows on a complicated metro network from large scale data from the smart card system and (2) unusual phenomena and their propagation on a spatio-temporal space. Moreover, (3) we integrated two forms of big-data (data from the smart card system and Twitter) into a visual exploration system to explore causes and/or effects of unusual phenomena. The case studies showed the possibilities and usefulness of our environment to observe real situations during the events. We plan to provide mechanisms for automatic event detection, prediction, and visualization through fusing various kinds of big data streams and prediction of passenger flows on wide and complex transportation networks.

References

[1] E. R. Tufte, *Beautiful Evidence*. Graphics Press LLC, 2006.

[2] M.-P. Pelletier, M. Trépanier, and C. Morency, "Smart Card Data Use in Public Transit: A Literature Review," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 4, pp. 557–568, 2011.

[3] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban Computing: Concepts, Methodologies, and Applications," *ACM Trans. on Intelligent Systems and Technology*, 2014.

[4] I. Ceapa, C. Smith, and L. Capra, "Avoiding the Crowds: Understanding Tube Station Congestion Patterns from Trip Data," in *Proc. UrbComp'12*, 2012, pp. 134–141.

[5] L. Sun, D.-H. Lee, A. Erath, and X. Huang, "Using Smart Card Data to Extract Passenger's Spatio-temporal Density and Train's Trajectory of MRT System," in *Proc. UrbComp'12*, 2012, pp. 142–148.

[6] C. Tominski, P. Schulze-Wollgast, and H. Schumann, "3D Information Visualization for Time Dependent Data on Maps," in *Proc. IV'05*, 2005, pp. 175–181.

[7] S. Thakur and A. J. Hanson, "A 3D Visualization of Multiple Time Series on Maps," in *Proc. IV'10*, 2010, pp. 336–343.

[8] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva, "Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 2149–2158, 2013.

[9] Z. Wang, M. Lu, X. Yuan, J. Zhang, and H. van de Wetering, "Visual Traffic Jam Analysis Based on Trajectory Data," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 2159–2168, 2013.

[10] G. Andrienko, N. Andrienko, C. Hurter, S. Rinzivillo, and S. Wrobel, "From movement tracks through events to places: Extracting and characterizing significant places from mobility data," in *Proc. VAST '11*, 2011, pp. 161–170.

[11] G. L. Andrienko, N. V. Andrienko, P. Bak, D. A. Keim, and S. Wrobel, *Visual Analytics of Movement*. Springer, 2013.

[12] C. Tominski, H. Schumann, G. Andrienko, and N. Andrienko, "Stacking-Based Visualization of Trajectory Attribute Data," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 12, pp. 2565–2574, 2012.

[13] T. Cheng, G. Tanaksaranond, C. Brunsdon, and J. Haworth, "Exploratory Visualisation of Congestion Evolutions on Urban Transport Networks," *Transportation Research Part C: Emerging Technologies*, vol. 36, no. 0, pp. 296 – 306, 2013.

[14] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou, "LeadLine: Interactive Visual Analysis of Text Data through Event Identification and Exploration," in *IEEE VAST*, 2012, pp. 93–102.

[15] A. M. MacEachren, A. R. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford, "SensePlace2: GeoTwitter Analytics Support for Situational Awareness," in *IEEE VAST*, 2011, pp. 181–190.

[16] B. Pan, Y. Zheng, D. Wilkie, and C. Shahabi, "Crowd Sensing of Traffic Anomalies based on Human Mobility and Social Media," in *SIGSPATIAL/GIS*, 2013, pp. 334–343.

[17] D. Yokoyama, M. Itoh, M. Toyoda, Y. Tomita, S. Kawamura, and M. Kitsuregawa, "A Framework for Large-Scale Train Trip Record Analysis and Its Application to Passengers' Flow Prediction after Train Accidents," in *PAKDD (1)*, 2014, pp. 533–544.

[18] T. N. Dang, L. Wilkinson, and A. Anand, "Stacking Graphic Elements to Avoid Over-Plotting," *IEEE Trans. Vis. Comput. Graph.*, vol. 16, no. 6, pp. 1044–1052, 2010.

[19] N. V. Andrienko and G. L. Andrienko, "Spatial Generalization and Aggregation of Massive Movement Data," *IEEE Trans. Vis. Comput. Graph.*, vol. 2, no. 17, pp. 205–219, 2011.

[20] M. Sarkar and M. H. Brown, "Graphical Fisheye Views," *Commun. ACM*, vol. 37, no. 12, pp. 73–83, 1994.

[21] T. M. J. Fruchterman and E. M. Reingold, "Graph Drawing by Force-Directed Placement," *Software Practice and Experience*, vol. 21, no. 11, pp. 1129–1164, 1991.