Journal of Visual Languages and Computing xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

journal homepage: www.elsevier.com/locate/jvlc

Uncertainty-aware visual analytics for exploring human behaviors from heterogeneous spatial temporal data

Siming Chen[a], Zuchao Wang[b], Jie Liang[c], Xiaoru Yuan*,[a]

[a] Key Laboratory of Machine Perception (Ministry of Education), and School of EECS, Peking University, China
[b] Qihoo 360 Technology Co. Ltd., China
[c] Faculty of Engineer and Information Technology, The University of Technology, Sydney

ARTICLE INFO

Keywords:
Spatial-temporal data
Uncertainty
Visual analytics

MSC:
00-01
99-00

ABSTRACT

When analyzing human behaviors, we need to construct the human behaviors from multiple sources of data, e.g. trajectory data, transaction data, identity data, etc. The problems we're facing are the data conflicts, different resolution, missing and conflicting data, which together lead to the uncertainty in the spatial temporal data. Such uncertainty in data leads to difficulties and even failure in the visual analytics task for analyzing people behavior, pattern and outliers. However, traditional automatic methods can not solve the problems in such complex scenario, where the uncertain and conflicting patterns are not well-defined. To solve the problems, we proposed a semi-automatic approach, for users to solve the conflicts and identify the uncertainties. To be general, we summarized five types of uncertainties and solutions to conduct the tasks of behavior analysis. Combined with the uncertainty-aware methods, we proposed a visual analytics system to analyze human behaviors, detect patterns and find outliers. Case studies from the IEEE VAST Challenge 2014 dataset confirm the effectiveness of our approach.

1. Introduction

Data recording human behavior becomes more and more in volume and diversity. With the development of the techniques, the GPS can record people's position and movement, the transaction system in the bank can record people's purchase and billing behavior, and more social media data would reflect people's attitude towards public affairs, or even eating preference. Facing the heterogeneous data, we can adopt visual analytics to understand the people behavior, find patterns and detect outlier events.

Directly using heterogeneous data analyzing process could lead to difficulties and even failure for the visual analytics tasks. This is because the data are often heterogeneous and imperfect. There can be various uncertainties in the data, including errors, data missing and conflicts. The data can also be in different resolutions. However, traditional automatic methods can't solve the problems in such complex scenario, where the uncertain and conflicting patterns are not well-defined. Our approach combines both algorithmic methods together with interactions in visualization, to enable users to identify, mark and refine such uncertainty issues. Together with such uncertainty-aware methods, we proposed a visual analytics system for supporting human's spatial temporal behavior analysis from the heterogeneous data.

In this paper, we report different kinds of uncertainties that we identified in a visual spatial data analysis and demonstrate how we refine them with the semi-automatic methods. Generally, our methods are data-driven reliability improvement methods. For different types of data, we have proposed different solutions and adopt cross referent of multiple sources of data. As there are heterogeneous data sharing the same attributes, but with different granularity, we can get finer resolution data with uncertainty from other types of data. With these approaches, we can better understand people's behavior in different dimensions and mark the reliability for further analysis. Uncertainty identification and analysis vary much and are challenging to solve through pure computation methods. So in our work, combined with visual identification and automatic preprocess methods, our methods have users in the visual analytics loop. Thus, users can better explore the different reliability of the data and further analyze outlier events.

Throughout this work, we use the fictitious datasets from IEEE VAST Challenge 2014 Mini Challenge 2 [1]. Combined with the uncertainty-aware approach, our proposed visual analytics system is able to summarize the general movement patterns of a group of people, and help analysts detect abnormal events, with various visualization view and multiple filters. In summary, our contribution is as follows.

* Corresponding author.
E-mail addresses: csm@pku.edu.cn (S. Chen), xiaoru.yuan@pku.edu.cn (X. Yuan).

https://doi.org/10.1016/j.jvlc.2018.06.007
Received 19 September 2016; Received in revised form 12 June 2017; Accepted 26 June 2018

- **Semi-automatic Uncertainty Refinement Methods.** We summarized five general types of uncertainty and proposed novel solutions for each. To solve the ill-defined uncertainty problems, we combine users' capability and algorithmic methods and allow human in the analysis loop.
- **Uncertainty-aware Visual Analytics System.** We have developed a comprehensive visual analytics system, incorporating the uncertainty-aware approaches and multiple coordinated visualization views, thus providing a full solution for understanding the human behaviors and detect interesting patterns and outliers.

This paper is structured as follows. Section 2 reviews related work. After introducing the data in Section 3, we present the uncertainty summary and general description of solutions in Section 4. We present the details of uncertainty-aware approach in Section 5. We present the visual analytics procedure and technical details in Sections 6 and 7. We demonstrate the use of our tools in four case studies. Finally, we discuss the limitations, future work, and conclusion.

2. Related work

Behavior analysis usually focuses on pattern extraction [2], relationship identification [3] and people clustering [4]. In People Garden [5], Xiong et al. summarized the temporal behavior of each person with a flower metaphor, and put them into different categories. Kanda et al. [3] analyzed the movement of museum visitors, focusing on hotspots and different visiting strategies. Orellana et al. [6] studied the interactions of people with a mobile game dataset. User behavior data usually involve spatial/temporal dimensions. Space-time cube [7,8] is a basic, yet intuitive 3D visualization metaphor. To encode multivariate data, informative glyphs are used in the spatial temporal scene [9]. To further facilitate exploration and analysis, filtering [10], clustering [11] and aggregation [12] methods are applied to spatial temporal data. VIS-STAMP [13] involves techniques such as linked parallel coordinates, SOM to analyze multivariate information of spatial temporal data. However, the previous work in spatial temporal visual analytics mostly work on regular dense-sampled GPS data. We provide the spatial temporal aggregation and filtering techniques, more importantly, we address on the heterogeneous spatial temporal data, which inherently have the conflicts and uncertainty.

User behavior analysis is also an important topic in intelligence analysis. Gorg et al. proposed a series of work on intelligence analysis with the aid of visual analytics [14]. Pirolli and Card [15] summarized the cognitive tasks in intelligence analysis. Analysis goal is to understand the semantics behind the data. In order to reveal the semantics of detected behaviors, many researches resort to map data, Point-of-Interests (POIs) data and social network data [16,17]. For example, Krüger et al. [2] visualized the Twitter keywords around trip destinations. In this way, they try to explain why people make these trips. Our work is heavily based on fusing different dataset, and we have to handle the uncertainty issues for a confident pattern analysis and event detection.

Correa et al. summarized a general framework for uncertainty-aware visual analytics [18]. They integrated the uncertainty information in the classical information visualization pipeline, and summarized the uncertainties in each stages. Following this mind, Wu et al. investigated how uncertainties propagate in such pipeline and proposed a flow-based uncertainty model to illustrate the propagated uncertainties from each analytical step [19]. In recent years, MacEachren proposed a new perspective that analysts should reason under uncertainties [20]. These are high-level summarized from the analytical pipelines. In our work, we proposed a visual analytics system that integrates algorithmic and interactive methods to deal with the uncertain data from heterogeneous data. To our knowledge, prior works did not discuss from such perspectives yet.

In behavior analysis, the data are usually imperfect and contains a lot of uncertainties. Various errors, data missing and conflicts exist in the data, which should be handled properly before any analysis can be performed. Sacha et al. summarized the visual analytics taxonomy by addressing the uncertainty processing in the loop [21]. Mac et. al summarized different techniques for visualizing spatial uncertainties [22]. Gschwandtnei et al. summarized the visual representation of temporal uncertainty [23]. Besides the visual representation, there are several works aiming to deal with uncertainty with visual analytics. Kang et. al proposed a visual analysis technique [24] to resolve the uncertainties caused by entity duplication. More recently, Slingsby et al. [25] studied the uncertainty of multi-attribute census classification. Lu et al. [26] visualized the uncertainties in map matching. Liu et al. proposed an uncertainty-aware method in the social network visual analytics work, which successfully identify and visualize the uncertainties. Our former work addressed on the trajectory uncertainty derived from geo-tagged social media data [27]. However, above works mainly focus on resolving one or two types of uncertainties. In this paper, we report an empirical study of the uncertainties in an intelligence visualization scenario. We summarized the uncertainties into five categories and proposed visual analysis methods to handle each of them.

3. Uncertainty taxonomy

In the paper scope, we mainly discuss the uncertainty in spatial temporal data for exploring human behaviors. The data unit representing human behavior is defined as event. The event is defined with the following attributes - time, location, people. Thus, these attributes and event are the targeting objects for analyzing uncertainty. For each attribute, we summarize the following uncertainty types,

Objects\Uncertainty\Examples	Missing Information	Confliction	Granularity	Multiple Values	Errors
Time	Missing Time records	Two referring time for the same event	Day/hour/minute	Multiple time slots for the same event	Time Error
Location	Missing Location Records	One object appearing in two locations at the same time	Lat/Long, Store/street/city name, JPG file	Multiple objects in the same locations	Systematic shift and GPS error
People	Missing identities (e.g. name, id)	Conflicting identities	Clear and vague identities	Ambiguous identities	Records errors
Event	Missing important descriptions	Conflicting descriptions of time, location and people	Multiple level of details description	Multiple descriptions of time, location and people	Description errors

Fig. 1. Uncertainty taxonomy. We discuss the uncertainty within two dimensions, including objects and uncertainty types. We consider time, location, people and event as objects, while we summarize the five related uncertainty types in exploring user behaviors with heterogeneous spatial temporal data.

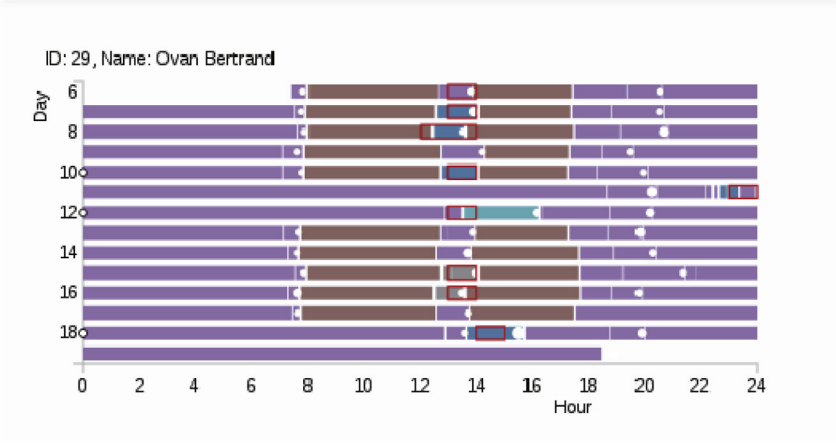

Fig. 2. Event timeline view, visualizing the enriched data events, including visiting POIs, staying time and transaction events each day. Periodicity is well represented in the view.

including missing information, conflict, granularity issues, multiple values and errors. We propose a taxonomy of uncertainty, as the basis for dealing with these uncertainties (Fig. 1).

The five types of uncertainties are derived from the analysis of heterogeneous spatial temporal data. First, missing information directly cause the lack of information to identify the objects. Second, conflict indicates there are conflicting descriptions stored in the heterogeneous dataset to represent the same identified object. For example, we might find the situation that the same person appeared at two different locations at the same time. Such uncertainty is caused by the conflicts of data. Third, the granularity issue in uncertainty is that the resolutions of descriptions for objects from dataset are different. For one event, we might have the day-level and second-level description at the same time. Fourth, multiple values lead to uncertainty because of lacking information to differentiate values. For example, in one location, there are multiple stores. From the specific location, it's hard to identify the exact stores only based on the spatial information. Lastly, the errors reduce the trust of the data and lead to uncertainty. For example, the GPS trajectories log might be error because of the transmission, encoding and decoding process of the records. With the five types of uncertainty for four objects, we illustrate the representative uncertainty with application data and the corresponding solutions in the following sections.

4. Uncertainty illustration

In this section, we first describe the data we used. After that, we introduce the data fusion method and visual analysis system, which is the basis for the uncertainty processing and classification.

4.1. Data description

Throughout this work, we use the fictitious datasets from IEEE VAST Challenge 2014 Mini Challenge 2 as example data. All datasets are related to a country, Kronos. In the capital, Abila, a big company called GAStech experienced a kidnap. It is suspected that some employees assisted the kidnap, therefore the GPS logs of their cars are provided. The ownership of each car is recorded in a car assignment file. Besides, the transaction logs of the employees are provided, as well as a raster format tourist map of Abila, a vector format road network and a name list of the employees.

The GPS log and transaction datasets cover 54 employees and a time span of two weeks, Jan. 6–19. There are 685,171 GPS records at one-second resolution. There are two transaction datasets: a loyalty card dataset with 1391 records, and a credit card dataset with 1491 records. Most transactions are recorded in both datasets, but they have different temporal resolutions. The credit card dataset is one-minute level while the loyalty card only records the date.

4.2. Data fusion

Our major tasks are as follows: 1) Describe the general daily life pattern of GAStech employees; 2) Detect the abnormal events or patterns among the employees. We combine different datasets with the concept of movement event. The events are first defined from the GPS logs, as stops above 1 min. Each event is naturally associated with a car, a time span, and a location. In most cases, an employee has a car, recorded in the car assignment data.

After that, we enrich the event data with POIs and transaction information. For POI enrichment, we first manually extract the 41 public POIs from the tourist map and put them into 10 categories. Additionally, we try to identify the home of each employee as the most frequently visited location at 4:00 am. This is treated as a special POI. Then for each event, if its location is within the boundary of a POI, it is given a corresponding label, e.g. "GAStech", "restaurant", "shop", "home". Otherwise, it is given a "non POI" label. For transaction enrichment, we first merge the credit card records, loyalty card records and membership card records. If the transaction time is within the time span of an event, it is assigned to that event.

The fused event data can be visualized with an event timeline (Fig. 2). The timeline can summarize the movement events of one person. The X axis represents an hour of a day while the Y axis represents each day. Each event is represented as a rectangle, with the color showing POI categories, and position showing its start and end times. The white circle indicates the transaction record, the size of which is the price paid. The red rectangle indicates the result of outliers detection methods, to help people analyze the data.

4.3. Representative uncertainty types

Based on the event definition, we can build up the visual analytics system to derive interesting patterns and detect the abnormal event. To fully support the analytical tasks, we identify five representative types of uncertainty from our taxonomy.

- **POI Uncertainty.** For some scenarios, we don't have digitalized map with exact positions. To convert the POI (Point of Interest) from map to exact position with latitude and longitude information, there are uncertainties in this process.
- **Temporal Uncertainty** There are different granularity from different sources of temporal event data, e.g. transaction and membership card. Moreover, there are conflicts or error time records for

the same event.

- **Transaction Attribute Uncertainty** The value of attributes, such as money, would have conflicts in two sources of bills. And there are some purchases without the price records.
- **Location Uncertainty** For spatial location, sometimes the GPS logs and credit card records indicate the same person at different locations at the same time. The GPS logs may have data missing, signal shift and noise.
- **Identity Uncertainty** For the people information, there are some records missing the identities. For example, the car assignment data is incomplete without the driver's name.

Based on these observations, we proposed a semi-automatic uncertainty process methods. We provide users visual results and allow users to control the automatic processing input as feedback and finally gain the reliable results.

5. Semi-automatic uncertainty processing

Generally, we have three types of operation guideline dealing with different types of data uncertainty. First, we mark and differentiate the missing data for each data sources, including GPS log missing, transaction data missing. The marking process is based on the understanding of the data distribution based on several reasonable assumptions in human behavior. Second, for data expressing the same events from different sources, we would refine the data with higher resolution from others. Specifically, we match the events from the low temporal resolution membership card to the debt/credit card records with higher resolution. Thus, we can generate an enriched transaction data source, with location, time, price, membership from multiple sources. Lastly, we correlate multiple data sources with sharing attributes and find the conflicts. For the debt/credit card uncertainty, we would mark the transaction missing and price conflict referred from membership card. We also correlate the transaction data and trajectory data, finding the transaction time error and uncertainties and visiting location uncertainties.

In all, it is a data-driven methods, incorporating the algorithmic and interactive operations for uncertainty refinement. In the following part, we report in detail how we deal with each category of uncertainty (Fig. 3).

5.1. POI uncertainty

In some scenarios, we can only get the scanned version of map and route information, which doesn't include the exact geographic information of buildings. It introduces challenges in the accurate analysis of the spatial temporal data. In such situation, the POI information is not given directly. What we have is a tourist map in JPG format. Although important POIs are marked on the map with specially designed icons, their spatial boundaries are highly inaccurate. In order to get the boundaries with higher possibilities and minimize the uncertainties, we run a two-stage POI-extraction methods with both the help of visual interface and geographic matching mechanism. We first extract the POIs from the map, each with a rough boundary. Then we use GPS and transaction log information as the reference to refine the boundaries (Fig. 4).

5.1.1. POI information extraction

We load the map and route data and registered them together in the system, which can provide hints for the positioning. With a map-based interface, we interactively add a rough boundary for each important POI. This boundary is set based on the POI icon and road network. We also try to avoid the overlap between different POIs. We summarized the POIs to different function categories. The categories include "GAStech" (work place), "restaurant", "shop", "home", etc. Then we assign a color for each POI category.

5.1.2. POI boundary refinement

In the second stage, we proposed a data-driven method to refine the POI boundaries. We have a hypothesis that if many people go to a place, then it should be in a POI region. Specifically, we use GPS data and transaction logs to refine the POI boundaries. After plotting the POI boundaries and the trajectories on the map, we found several people gathering in the "Non-POI" region. However, there's usually a POI nearby, and very likely these people are actually staying in this POI. In such cases, we expand the boundary of nearest POI to cover this region. Sometimes there are transaction logs to cross-verify the marking results. With the shop name in the transaction log, we can know for sure that they are visiting a certain POI. The detailed algorithm is described in Algorithm 1.

There are two thresholds settings in Algorithm 1. We allow POIs shifting when most people are out of the original POI but near the boundaries. So we tend to restrict the percentage higher as larger than 60 percentage after several rounds testing. Because if there are no strong signals all people are staying outside, we would keep the original shape and position of the POI. The other one is distance of POI shifts. Considering the facts, we use one street blocks distance as the threshold.

However, we are aware the situation that there could be parking slot where is in the shoping places. Current algorithm can not deal with it, we need users to manually identify the parking places boundaries.

Attribute Space	Dealing with Uncertainty	Map	Trajectory Data	Transaction data (low resolution)	Transaction data (high resolution)	Description
POI information	POI region range extraction	√				Extract a range with uncertainty from the JPEG map
	POI region range refinement	√	√			Based on common pattern extracted from the trajectory data
Temporal information	Resolution refinement			√	√	Mark the detail time based on the matched event with high resolution
	Transaction time error and conflict		√	√	√	Mark the conflict time for each transaction events
Transaction attribute	Data missing			√	√	Add and mark the missing records
	Price conflict			√	√	Mark two values
Location information	Location conflict		√	√	√	Mark the conflict location
	Location missing, shift and noise	√	√	√	√	Mark the 'jump' position and shift based on common pattern and POI, using the transaction for verification
People information	Missing car assignment		√	√	√	Find people who have the matched transaction and location of the car

Fig. 3. Classification of uncertainty and work process. We consider the analyzing attributes based on usually used data source for understanding people's behavior, including map, GPS trajectory data, multi-level temporal event data, etc.

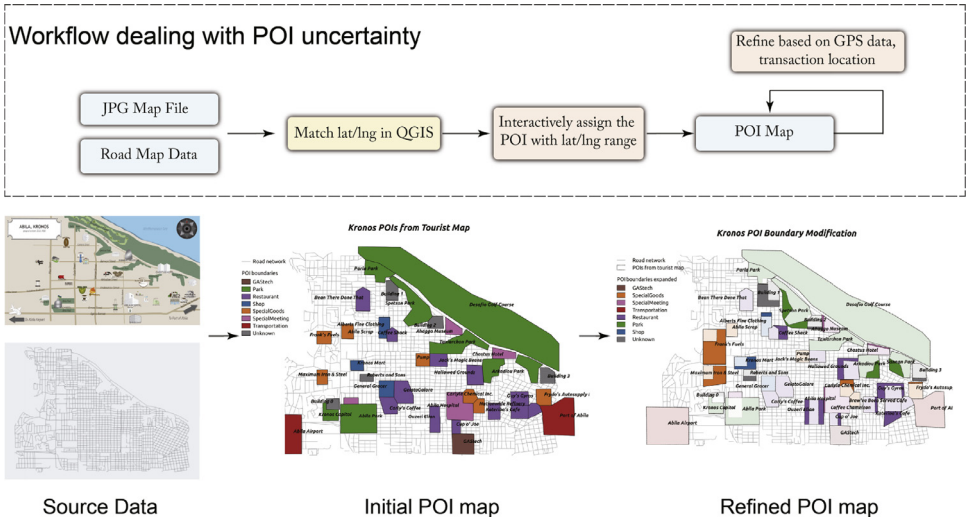

Fig. 4. Workflow Dealing with POI uncertainty. It includes two main steps, POI information extraction and POI boundary refinement.

5.2. Temporal uncertainty

Uncertainty in temporal information would affect the judgment on the events, which would lead to inaccurate or wrong findings if we used an event with uncertain time as an evidence or other analysis resources. We have temporal information from GPS logs and transaction data. This information could be in different temporal resolutions, and can have conflicts. We deal with them as follows.

5.2.1. Temporal resolution refinement

Two different types of transaction data can be in different temporal resolutions. In our case, the resolution for our credit card is one minute, and for our loyalty card data is one day. The one-day resolution for loyalty card data resolution is too low to support further behavior analysis. Therefore, we correlate it with multiple sources of transaction data to improve the resolution. The resolution refinement process is based on transaction events matching. Our assumption is that most events are recorded in both dataset. Therefore we match the transactions in the same day with identical location, price, person. Then we set the corresponding high resolution time stamp to the low resolution time stamp. In some situation, there would be one-to-many matches. We will do the refinement with multiple candidates and mark it as uncertain in location or price.

5.2.2. Temporal error and mismatch

To detect the temporal error, we first process the transaction data and generate its temporal distribution, from which we can see several outliers. The unreasonable distribution indicates possible temporal errors. For example, we find several people with transaction records at exactly 12:00 every day. As shown in Fig. 5-bottomleft, while the credit card data says they are in some restaurant, the GPS logs show that they are not. However, these people usually have been to that restaurant in the morning according to the GPS information, but without transaction records at that time. From the conflicts shown in the visualization, users can conclude that there are temporal errors in the credit card data. To fix such error and get the correct transaction time, we use the time indicated by the GPS logs, and associate the transaction with the movement event in the morning (Fig. 5-bottomright). We summarize

Input:
a list of *stay_events* E_i;
a list of POIs P_j;
Output:
refined POIs;

Extract *non_poi_events* from *stay_events*;
Count = 0;
for each *stay* in *non_poi_events* **do**
  for each *POI* in *POIs* **do**
    $< poi, distance >= calculate\_POI\_Distance(stay, POI)$
  end for
  $nearest\_poi = calculate\_nearest\_POI(< poi, distance >)$;
  $stay.potentialPOI.push(nearest\_poi)$
end for
Merge the duplicated POI in each *stay.potentialPOI* from *non_poi_events*, and add with count
for each poi in *merged_poi_list* **do**
  if poi.count > *threshold_count* and poi.distance < *threshold_distance* **then**
    Extend the poi with distance, and mark refinement
  end if
end for

Algorithm 1. POI Range Refinement.

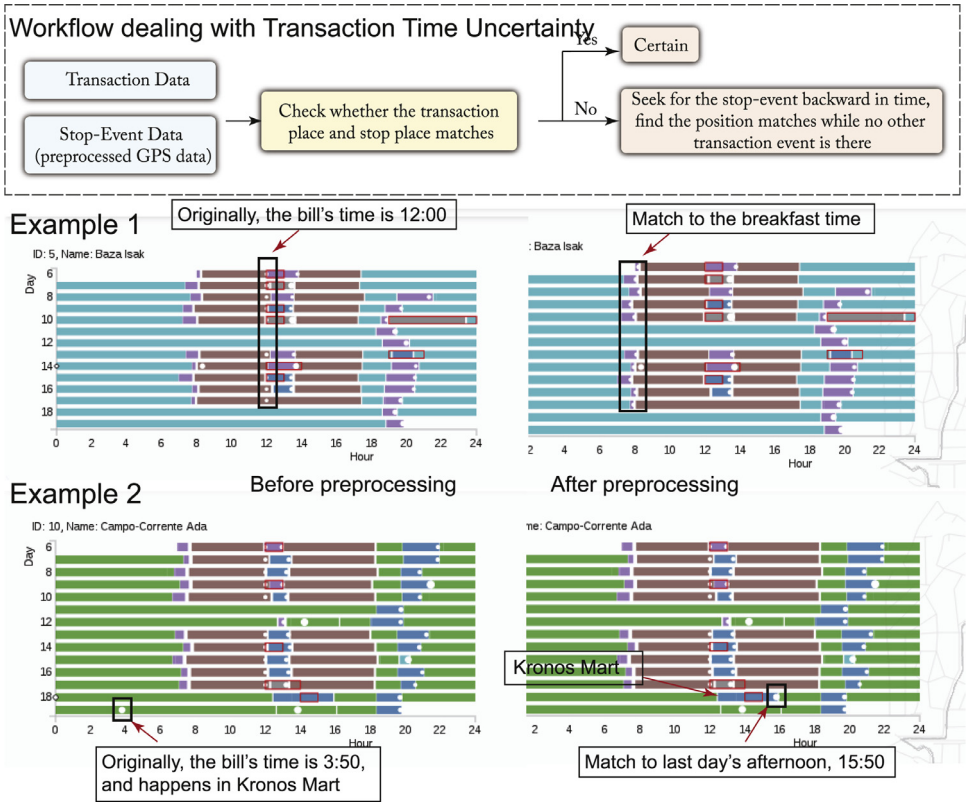

Fig. 5. Workflow dealing with transaction time uncertainty. By identifying the error, different granularity and conflict in the temporal event data, we are able to refine and mark the data with different reliability.

the general procedure to deal with such errors in Fig. 5-top.

Based on the users' perception and judgment on the visualization, we proposed a time matching algorithm. For each suspected time value in the transaction dataset, we trace the spatial visiting events of GPS data in the nearby time range. With the cross verification of different data, we can automatically fix the visiting and payment time mismatch (Algorithm 2). In some scenarios, we can't find matching location and time, we show two possibilities for visiting time and mark the data uncertain.

5.3. Transaction attribute uncertainty

In multiple sources of transaction data, each data source has shared and unique attributes. The shared attributes in credit card data and loyalty card data, e.g. the price value of the transaction, the name of the store, might conflict. Also, there might be data missing. In such situation, when we have no ground truth, we mark both possibilities. For such situation, we enrich the transaction data set by merging different sources of data, in awareness of the uncertain situation (Algorithm 3).

5.4. Location uncertainty

There are several types of uncertainty for location information, including data missing, data shift, and noise.

5.4.1. Data missing

GPS data can have considerable data missing due to signal loss, GPS device off, etc. In such case, we can see vehicles jumping to a position

Input:
a list of *outlier_times*;
a list of *transaction_events*, ordered by time;
a list of *stop_events*, ordered by time;
Output:
transaction_events_fixed_time;
for each *event* in *transaction_events* **do**
  if *event.time* is in *outlier_times* **then**
    Get the *stop_event* which includes the *event.time* for this person's *stop_events*;
    Trace back each *stop_event* of the *stop_events*, starting from current *stop_event*
    if *stop_event.location* == *event.location* and no other transaction event during *stop_event.time* **then**
      Fix the *event* time to *stop_event.endtime*.
    end if
  end if
end for

Algorithm 2. Time Error Matching.

Input:
*transaction_data*1;
*transaction_data*2;
Output:
enriched_transaction_data;
Sort the *transaction_data*1 and *transaction_data*2 by time;
merged_transaction = *merge*(*transaction_data*1, *transaction_data*2);
for each *event* in *merged_transaction* **do**
  if not (*event.time* in *transaction_data*1 and *event.time* in *transaction_data*2) **then**
    Mark the event < *data_source*, *uncertainty_in_missing* >; //merge and mark the uncertainty
  else
    //event belong to both transaction
    if *price* conflicts in two transactions **then**
      Mark the event with two prices < *price*1, *price*2, *uncertainty_in_price* >;
    end if
  else
    Mark the event as *certainty*
  end if
end for

Algorithm 3. Transaction Data Merge.

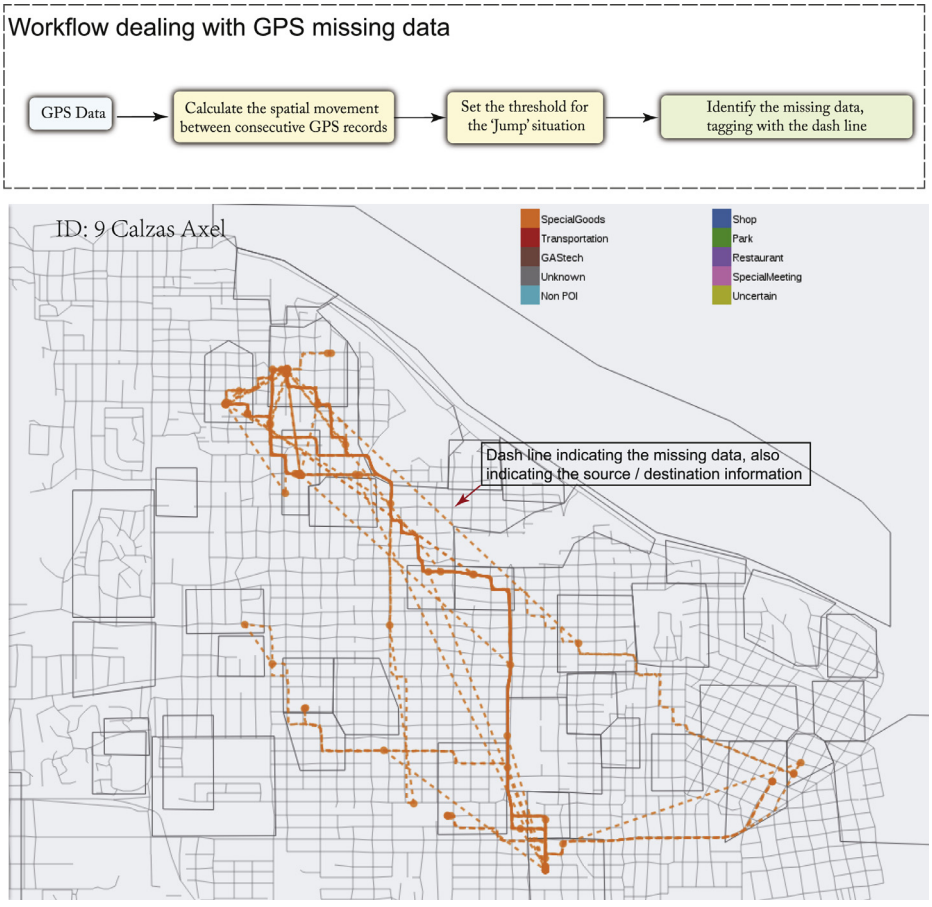

Fig. 6. Dealing with the uncertainty caused by GPS location data missing. By setting a threshold of speed, we can differentiate the regular sampling and GPS error sampling data.

far away while missing the intermediate routes. We calculate the speed and continuous GPS positions. If the speed or moving distance is unreasonable, we visualize them with dashed lines, as shown in Fig. 6. In this case, we use 100 km/h as the threshold to detect jumping points. The threshold is referred in our previous work [28], with the calculation of normal speed in the city. Based on this, we can make the later exploration stage have a better understanding of the data, and make a decision taking consideration of the uncertainty.

Input:
stop_events, calculated by the GPS data;
POIs;
Output:
generalpatternsofeachperson;
Separate the *stop_events* by weekday and weekends;
for each people's *stop_events* **do**
  $frequent\_visit$ = []
  for $i = 0; i < 24; i++$ **do**
    //calculate by hours
    POI = the place people stays longest in $hour_i$;
    $frequent\_visit$.push(*POI*)
  end for
  calculate the most frequent POI from $frequent\_visit$ for each hour.
end for

Algorithm 4. General Pattern Detection.

5.4.2. Location data shift and noise

GPS logs can have location shift and noise, but such cases are not easy to detect. Being unaware of the shift and noise would fail nearly all the analysis based on the people with such error data. In our approach, we first summarize the people's movement pattern, then find the outliers and finally shift back the error data based on the general patterns.

First, we calculate the general pattern for each people (Algorithm 4). We identified people's home place based on the place they stay at 4 a.m. The result is visualized in the Fig. 7. The x-axis is split by weekday and weekend, and each block represents one hour. The y-axis represents each people. Color indicates different types of locations. We can easily target to the regular pattern and also find out the outliers. Second, after we detect outliers from the visualization, we calculate the shift of frequent visiting places from the regular patterns and shift the uncertain data back. With a series of key positions on the map, we can calculate the shift distance (Algorithm 5). Lastly, we verify the correspondences between the shifted visiting places and the transaction location and make sure our shifting is correct. The threshold setting process and shifting judgement is highly involved with users.

As an illustrative example in Fig. 8-bottomleft, we have identified a person who spends the most time in an "Uncertain" POI (yellow), and never goes to the GAStech company (brown). This is strange. By plotting his trajectory on the digital map, as shown in Fig. 8-middle left, we find that his trajectory is highly noisy, and seems to have a location shift. By referring to other people's trajectories, we are able to identify the shift, and move the trajectory back to the correct location (Fig. 8-middleright). Now the POIs on the event timeline are much more reasonable. We find he now spends the working time at the GAStech company, and the locations in GPS logs matches perfectly with that in the transaction records (Fig. 8-bottomright). The noise is discovered visually, as shown in Fig. 8-middleleft. We downsample the data to reduce the noise and describe a general pattern within a longer time window.

5.5. Identity uncertainty

People records are not always perfect in the real case. In the example, the car assignment data is not complete. There are nine truck drivers without car assignment records, and five cars without driver information. To deal with the uncertainty, we refer to the GPS data and transaction logs. The GPS data records the movement of cars and transaction logs record the activity of persons. If they match, e.g. a car

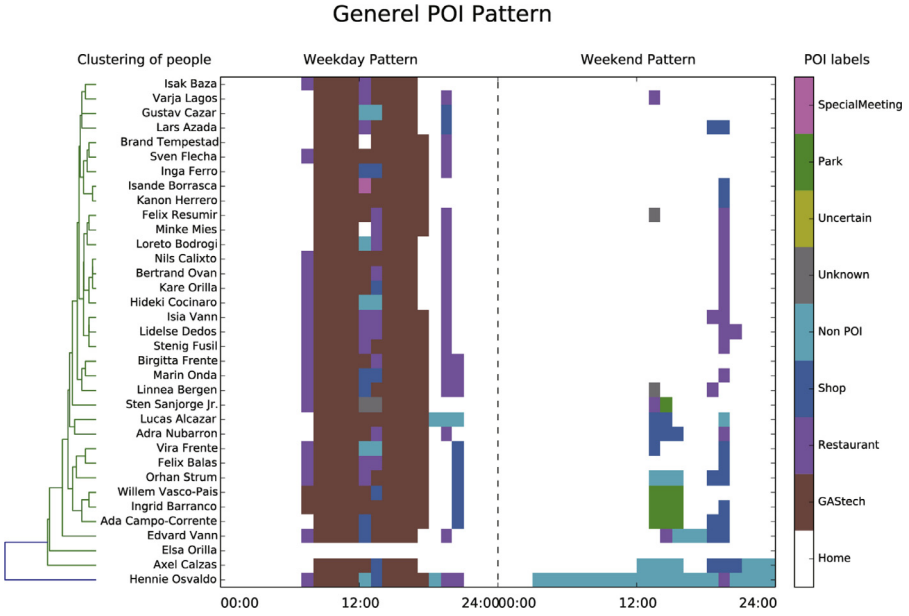

Fig. 7. General pattern detection of people in weekday and weekend. By identifying the dominating visiting places of each people, we also rank the people by their visiting similarity. Outlier can be easily detected.

Input:
outlier_people, targeted in the visualization;
frequent_visiting_POIS;
Output:
outlier_people_shifted;

```
1:  for each visiting in frequent_visiting_POIS do
2:      time = visiting.time;
3:      location = visiting.location
4:      for each stop_events in outlier_people's stop_events do
5:          if stop_events.time.hour == time.hour then
6:              distance = calculateDistance(stop_events.location, location)
7:              accumulateDistance.push(distance)
8:          end if
9:      end for
10: end for
11: Calculate the variance of accumulateDistance
12: if variance < threshold then
13:     Calculate the averageDistance of the accumulateDistance
14:     Shift all the data in outlier_people by averageDistance
15: end if
```

Algorithm 5. Shift Back the Uncertain Shifted Data.

goes to many locations and a person usually have bills there, we might think the car is assigned to that person. To perform the match, we use a trial and error methods. We visually map the stop event and transaction events in the event view (Fig. 9). Users are able to interactively select different people and different cars, and try to match them. In this way, we successfully find the car assignment for these nine drivers and five cars. What surprises us is that several truck drivers are actually sharing a truck.

5.6. Uncertainty process summary

In the uncertainty process, we deal with multiple data source's uncertainties. We also correlate different data sources, with visual and computational methods to mark and differentiate the uncertainty. This

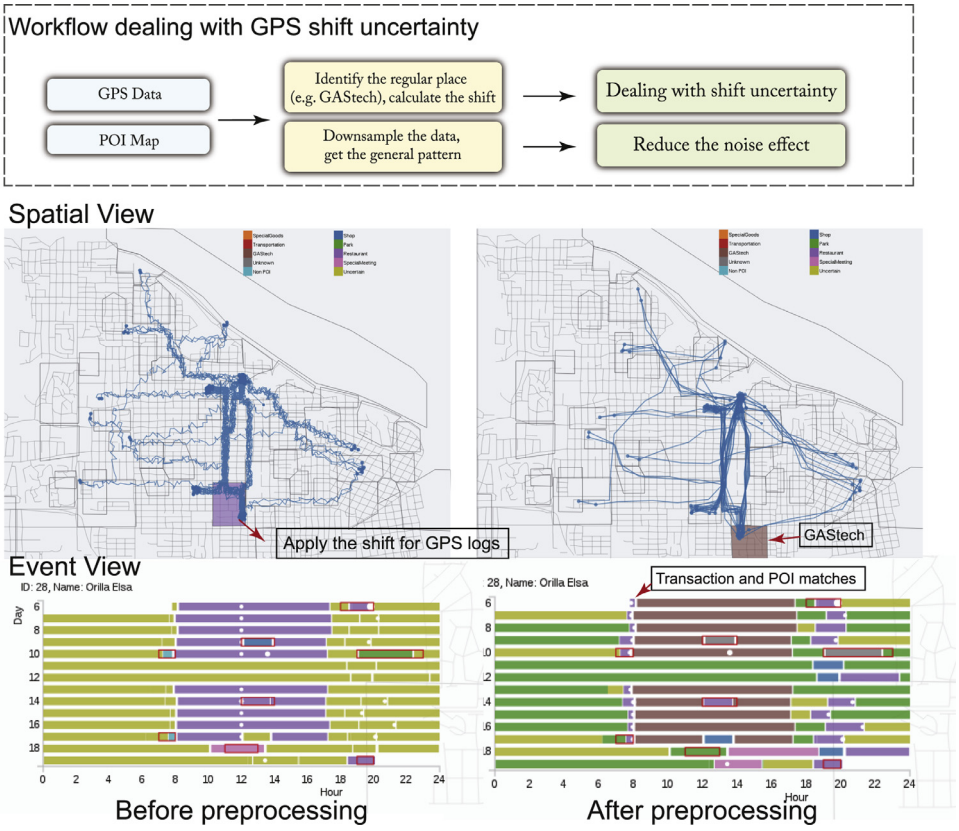

Fig. 8. Workflow Dealing with shift and noise in the location information. Shift operation is done based on general visiting pattern from all people and noise reduction is achieved by down sampling.

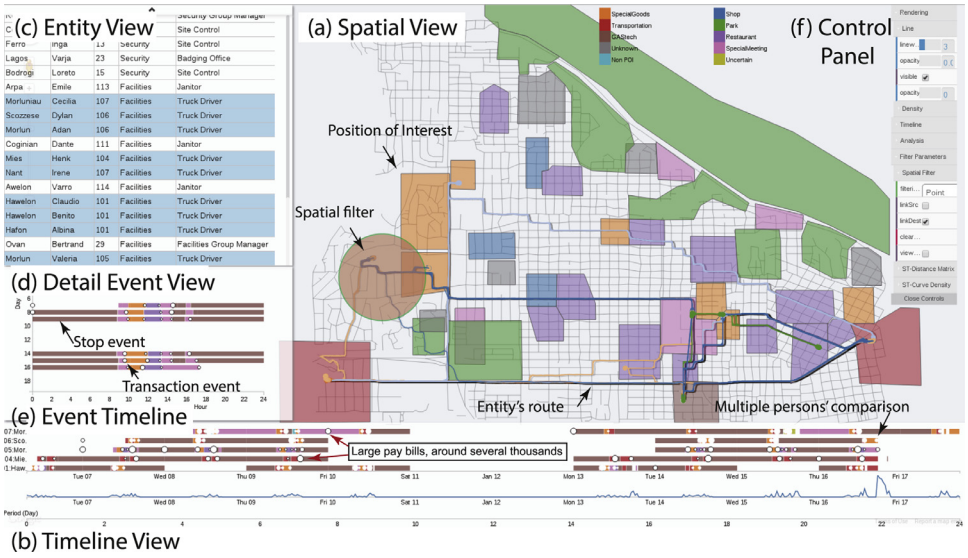

Fig. 9. Visual analytics system interface, which includes (a) Spatial View, supporting interactive filtering on the map; (b) Timeline View, showing aggregated movement logs number in each time bin; (c) Entity View, providing basic information of each people and apartment, which also supports multiple selection. (d) Detail Event View, providing detail behavior information for each people. (e) Event Timeline, dynamically comparing a group of people's behavior.

provides a solid and intuitive stage for later exploration. In the further analysis process, we need to consider the confidence level of the data, which directly affect the judgment and decisions.

6. Visual analytics system

Our visual analytics system combines the uncertainty-aware approaches with a fully interactive exploration functions. Our system can enable users to find reliable information, detect patterns and find outliers from the heterogeneous spatial temporal data sources (Fig. 9).

6.1. Spatial temporal exploration

Users can apply spatial temporal filtering to explore the data. **Map view** shows the positions of POIs and GPS tracks (Fig. 9a). Each POI is represented by a polygon, with color encoding the POI categories. Each GPS track is represented as a polyline. Users can apply spatial filters on the map to select GPS tracks passing single or multiple regions. **Timeline view** shows the temporal distribution of GPS records (Fig. 9b). Users can apply temporal filters on the timeline to select GPS tracks within single or multiple time ranges. In the exploration, users can filter the POIs within a range of time or time periods for further pattern analysis.

6.2. Pattern analysis and event detection

Entity view shows a name list of the employees (Fig. 9c). Users can directly select people on the list. **Detail event view** shows the whole event sequence of one employee (Fig. 9d). The function is already mentioned in the Data Description Section. Users can first analyze the basic daily patterns in the event sequence. Automatic outlier detection is provided for each individual movement, based on the derived regular patterns (Fig. 7). However, there are many false alarms from the automatic methods, since people would have special events such as going to the supermarkets or parks, etc. which are not necessarily suspicious events. Thus, we enable users to explore the people behaviors in the spatial, temporal and event view. Based on the outlier hints, users can find the suspicious events, including going out in mid-night, absence from work in working hours, card stolen event, etc. Furthermore, we support multiple persons' behavior comparisons for more complex pattern findings. **Event timeline** shows the event subsequence in the selected time range for multiple employees (Fig. 9e). It's mainly used to compare/correlate the behaviors of different people. Combined with other views, we find interesting behaviors such as get-together, car-sharing and other abnormal relationship between people, etc.

Based on the uncertainty-aware visual analytics system, users can find reliable patterns and events with interactive exploration.

7. System implementation

Our system is developed under a client-server architecture. The client is built with HTML5/Javascript, and the server-side services are implemented with Python and MongoDB. The client includes multiple web programming techniques and toolkits, such as Google Maps, d3 library, WebGL and Canvas. On the server side, we choose MongoDB to manage data sets because of its flexibility and scalability in handling multiple sources of data. The uncertainty processing part relies on both the visual interface and the python processing. We used the QGIS to identify and refine the boundary of JPEG map. We used python to generate hierarchical clustering results of people's patterns, which can help users identify the spatial uncertainty. For other situations, users identify visually and interactively in the system, and we can set the input for processing in the automatic methods. For the identifying process, we allow users to interactively match different sources of data to check whether the visiting place and transaction place is the same for one people.

8. Evaluation

We evaluated our proposed uncertainty-aware visual analytics methods in two aspects. First, we compare our methods with the pure computational methods and illustrate our advantages. Second, we use a case study to illustrate how users can find events successfully after dealing with the uncertainties.

8.1. Method comparison

We discussed the comparison part for our method with the computational uncertainty mining methods. At beginning, we use the pure automatic algorithms and found there are several problems. Especially, for some situations, the pure algorithm can not work out since it needs high levels of human judgment.

- **POI detection**. We compare our approach (Fig. 4 Step 1, 2) with the automatic matching approach (Fig. 4 Step 1). Our approach has done 16 more adjustments based on users judgment on the POI shifting. The total POI number is 41.
- **Temporal error and mismatching** The pure automatic method

does not work. Because the pattern is not well-defined, and the machine cannot understand the semantics of the visiting time. But the resolution improvement (Section 5.2.1) can be done with the time matching methods.

- **Transaction attributes missing and conflicts** The algorithm (Algorithm 3) can work and mark the conflict information, but we need users to finally judge the conflict information in the visual analytics process.
- **Location conflict, shift and errors** With the location matching algorithms (Fig. 7), we can find the general patterns of the representative visit at specific time for each person. However, the machine cannot tell the semantics of the shifting, so it finally detected with 276 potential uncertain shifts. However, finally we found there are one people with 24 shift events that are systematical shifting. It turns out pure automatic methods generate much more false alarms.
- **People information identity** With the visual interface, users can easily identify the match events that GPS and transaction logs sharing the same locations concurrently. We have not tested this part, but we can draw hypotheses that with human's perception, the identifying process could be efficient. However, we also note that the algorithmic calculation can search a large space with the advantage of computations.

In short, we summarized the informal evaluation of comparison; we can draw the following conclusions. Our method can equally works well for the pure computational tasks with the existing methods. Moreover, our method has advantage in reducing false alarm and improving efficiency over the existing method. In the tasks requiring semantic understanding, our method works while there are no existing works solving it.

8.2. Case study - people behavior analysis

The general exploration process would be illustrated by the case. In the IEEE VAST Challenge 2014 Data, there are 54 people in the GAStech company. Each people's behavior might have a relationship with a kidnap or not. Detecting two or multiple people's behaviors is one of the important task issues because criticism might happen through group activities.

Dealing with the heterogeneous data with uncertainty, we enrich the event data, with GPS logs, transaction data, membership card data and POI data. The temporal resolution of the membership card is refined based on the fine resolution debt and loyalty card transaction data. Also, we mark the transaction event as uncertain if there are at least one type(s) of data missing. Then we visualize the GPS movement in the spatial view. We find Elsa Orilla's trajectory is with noise and seems shift, as the color indicated she spent most of her time in the 'Non POI' region and spent working time in a cafe (Fig. 8-left). Based on the hypotheses that she should also work in the company in the weekday, we calculate the shifting distance and shifted the data back (Fig. 8-right). We also adopt downsampling techniques to acquire the trend of the movement. Based on these operations, we can get more reasonable trajectories, which are cross verified from the transaction data. The transaction data indicates the transaction events matches most of the GPS positions in the place dimension.

Our system has the ability to detect several types of abnormal events, which is built upon the basis of uncertainty processing. In this case, we find Elsa Orilla and Kanon Herrero have a suspiciously close relationship (Fig. 10). They do a lot of things together. The transaction pattern is that Herrero pays the bill and Orilla provide the membership card every lunch (except one day, Jan.10). Besides, they have three kinds of trajectory patterns. The first pattern is that Herreron drives to the restaurants while Orilla's car is parked at GAStech. This situation happens every weekday except Jan.10 and Jan.14, Jan.17. The second-weekday pattern is that Orilla drives the car and Herrero parks his car at GAStech. The third pattern happens on weekends. They drive together to many places, such as shops, parks or museums. All the bill is paid by Herrero. One outlier event is that Orilla went to the museum again in Jan.19 and paid herself.

The processing of uncertainty acts an important role in this case. We can see Orilla Bea's original trajectories are with systematical shift and noises (Fig. 8). After uncertainty processing, we could find this case. Thus, it turns out our uncertainty process is efficient and essential for the spatial temporal behavior analysis.

9. Discussion

We propose an uncertainty-aware visual analytics approach to deal with multiple sources of spatial temporal data. With both interactive and algorithmic methods, users can identify and refine the data uncertainty, which is challenging to conduct due to the ill-defined uncertain patterns. Such process requires the semantic understanding. For example, the abnormal visiting pattern can be detected with a large amount of false alarms. One people might go to supermarket not that regular, which can be detected as abnormal behaviors. However, in the semantic level, going to supermarket is a normal behavior. Moreover, for scenarios with multiple data sources and attributes, the algorithms can not easily find the accurate solution, which requires the involvement of human. In the complex data analysis scenario, the data-driven approach works better since there are no existing models for the analysis tasks.

Our visual analytics system was further evaluated through participating the IEEE VAST Challenge 2014. Our solution has found the most important patterns and outlier events in the contest and awarded as "the Most Comprehensive Visual Analytics System". The uncertainty processing is also highly evaluated. Compared with the ground truth, our approach identify all the uncertainty, error and conflicts in the dataset. Based on this, we generalize and summarize the key techniques in this paper.

Though novel and powerful, current uncertainty-aware approach still has the limitations. Some parts are label itensive, e.g. label POI, examining the error time. We could improve the manual operation parts with more intuitive operations and automatic matching methods. It can further improve the efficiency of our methods. In the future, we also envision to test our techniques for different data sources. The scalability of the system should be further evaluated with a larger scale dataset. In the current stage, we have not done the formal user study yet. We envision to conduct a user study in the future.

The biggest lesson we learned is that we need to reason under uncertainty. We should not assume there are no ambiguities, errors or conflicts in the data. What we believe is that we should analyze the data being aware of the uncertainty. In the other aspect, in identifying uncertainty, we need to be aware of different types and note that the uncertainties would be propagated in the whole visual analytics pipeline.

10. Conclusion

In this paper, we present an uncertainty-aware visual analytics system to investigate human behaviors from heterogeneous spatial temporal data. We summarize five representative types of uncertainty and its refinement methodology. A data-driven approach is proposed and we make full use of humans' judgment through a visual interface. With cross verification from multiple sources, we can further improve the reliability of the refinement results. Based on the refinement results, we are able to identify the patterns and events for behavior analysis.

Acknowledgements

We thank the contribution of Zhenhuang Wang, Chenglong Wang, Zipeng Liu and Zhengjie Miao. This project is supported by the National

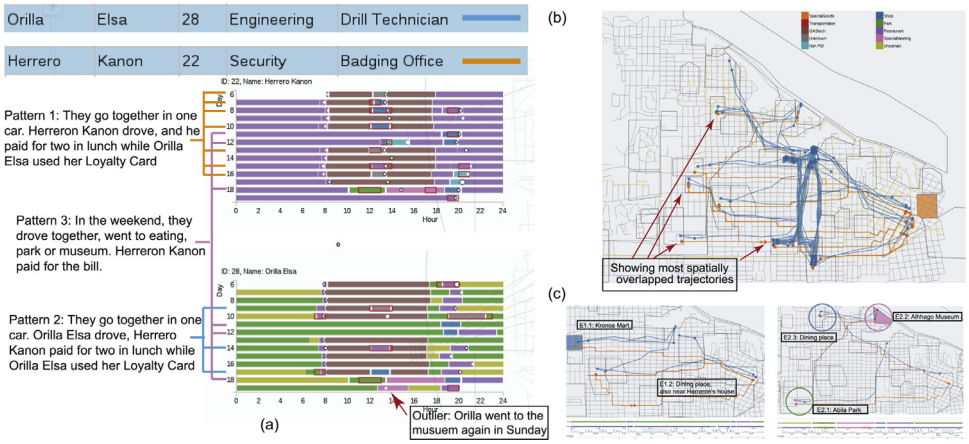

Fig. 10. Case Study of people behavior correlation. In the visual exploration we find two people have tightly correlated behaviors. (a) Coherent movement patterns in the event timeline. (b) Their sharing routes and POIs on the map. (c) Detailed view of special meeting places, including mart, dining places, park and museum.

Key Research and Development Program of China (2016QY02D0304), National Basic Research Program of China (973) (2015CB352503), and National Natural Science Foundation of China (61672055).

References

[1] K.A. Cook, G.G. Grinstein, M.A. Whiting, The VAST challenge: history, scope, and outcomes: an introduction to the special issue, Inf. Vis. 13 (4) (2014) 301–312, https://doi.org/10.1177/1473871613490678.
[2] R. Krüger, S. Lohmann, D. Thom, H. Bosch, T. Ertl, Using social media content in the visual analysis of movement data, Proc. Workshop on Interactive Visual Text Analytics, (2012).
[3] T. Kanda, M. Shiomi, L. Perrin, T. Nomura, H. Ishiguro, N. Hagita, Analysis of people trajectories with ubiquitous sensors in a science museum, Proc. of IEEE ICRA, (2007), pp. 4846–4853.
[4] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, W.-Y. Ma, Mining user similarity based on location history, Proc. ACM SIGSPATIAL GIS, (2008), pp. 34:1–34:10.
[5] R. Xiong, J. Donath, Peoplegarden: creating data portraits for users, Proc. of the ACM UIST, (1999), pp. 37–44.
[6] D. Orellana, M. Wachowicz, N. Andrienko, G. Andrienko, Uncovering interaction patterns in mobile outdoor gaming, Proc. GEOProcessing, (2009), pp. 177–182.
[7] T. Hägerstraand, What about people in regional science? Pap. Reg. Sci. 24 (1) (1970) 7–24.
[8] T. Kapler, W. Wright, Geotime information visualization, Inf. Vis. 4 (2) (2005) 136–146, https://doi.org/10.1057/palgrave.ivs.9500097.
[9] C. Tominski, H. Schumann, G.L. Andrienko, N.V. Andrienko, Stacking-based visualization of trajectory attribute data, IEEE Trans. Vis. Comput. Graph. 18 (12) (2012) 2565–2574, https://doi.org/10.1109/TVCG.2012.265.
[10] R. Krüger, D. Thom, M. Wörner, H. Bosch, T. Ertl, Trajectorylenses - a set-based filtering and exploration technique for long-term trajectory data, Comput. Graph. Forum 32 (3) (2013) 451–460, https://doi.org/10.1111/cgf.12132.
[11] G.L. Andrienko, N.V. Andrienko, S. Rinzivillo, M. Nanni, D. Pedreschi, F. Giannotti, Interactive visual clustering of large collections of trajectories, Proc. of the IEEE VAST, (2009), pp. 3–10, https://doi.org/10.1109/VAST.2009.5332584.
[12] G.L. Andrienko, N.V. Andrienko, Spatio-temporal aggregation for visual analysis of movements, Proc. of the IEEE VAST, (2008), pp. 51–58, https://doi.org/10.1109/VAST.2008.4677356.
[13] D. Guo, J. Chen, A.M. MacEachren, K. Liao, A visualization system for space-time and multivariate patterns (VIS-STAMP), IEEE Trans. Vis. Comput. Graph. 12 (6) (2006) 1461–1474, https://doi.org/10.1109/TVCG.2006.84.
[14] C. Gorg, Y. ah Kang, Z. Liu, J. Stasko, Visual analytics support for intelligence analysis, Computer 46 (7) (2013) 30–38, https://doi.org/10.1109/MC.2013.76.
[15] P. Pirolli, S. Card, The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis, Proc. of International Conference on Intelligence Analysis, (2005).
[16] N. Andrienko, G. Andrienko, G. Fuchs, Towards privacy-preserving semantic mobility analysis, Proc. IEEE EuroVA, (2013), pp. 19–23.
[17] R. Krüger, D. Thom, T. Ertl, Visual analysis of movement behavior using web data for context enrichment, Proc. IEEE PacificVis, (2014), pp. 193–200.
[18] C.D. Correa, Y.H. Chan, K.L. Ma, A framework for uncertainty-aware visual analytics, Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on, (2009), pp. 51–58, https://doi.org/10.1109/VAST.2009.5332611.
[19] Y. Wu, G.X. Yuan, K.L. Ma, Visualizing flow of uncertainty through analytical processes, IEEE Trans. Vis. Comput. Graph. 18 (12) (2012) 2526–2535, https://doi.org/10.1109/TVCG.2012.285.
[20] A.M. MacEachren, Visual analytics and uncertainty: its not about the data, in: E. Bertini, J.C. Roberts (Eds.), EuroVis Workshop on Visual Analytics (EuroVA), The Eurographics Association, 2015, , https://doi.org/10.2312/eurova.20151104.
[21] D. Sacha, H. Senaratne, B.C. Kwon, G. Ellis, D.A. Keim, The role of uncertainty, awareness, and trust in visual analytics, IEEE Trans. Vis. Comput. Graph. 22 (1) (2016) 240–249.
[22] A.M. MacEachren, A. Robinson, S. Hopper, S. Gardner, R. Murray, M. Gahegan, E. Hetzler, Visualizing geospatial information uncertainty: what we know and what we need to know, Cartogr. Geogr. Inf. Sci. 32 (3) (2005) 139–160, https://doi.org/10.1559/1523040054738936.
[23] T. Gschwandtnei, M. Bgl, P. Federico, S. Miksch, Visual encodings of temporal uncertainty: a comparative user study, IEEE Trans. Vis. Comput. Graph. 22 (1) (2016) 539–548, https://doi.org/10.1109/TVCG.2015.2467752.
[24] H. Kang, L. Getoor, B. Shneiderman, M. Bilgic, L. Licamele, Interactive entity resolution in relational data: a visual analytic tool and its evaluation, IEEE Trans. Vis. Comput. Graph. 14 (5) (2008) 999–1014.
[25] A. Slingsby, J. Dykes, J. Wood, Exploring uncertainty in geodemographics with interactive graphics, IEEE Trans. Vis. Comput. Graph. 17 (12) (2011) 2545–2554, https://doi.org/10.1109/TVCG.2011.197.
[26] L. Lu, N. Cao, S. Liu, L.M. Ni, X. Yuan, H. Qu, Visual analysis of uncertainty in trajectories, Proc. PAKDD, (2014), pp. 509–520.
[27] S. Chen, X. Yuan, Z. Wang, C. Guo, J. Liang, Z. Wang, X.L. Zhang, J. Zhang, Interactive visual discovering of movement patterns from sparsely sampled geotagged social media data, IEEE Trans. Vis. Comput. Graph. 22 (1) (2016) 270–279, https://doi.org/10.1109/TVCG.2015.2467619.
[28] Z. Wang, M. Lu, X. Yuan, J. Zhang, H.V.D. Wetering, Visual traffic jam analysis based on trajectory data, IEEE Trans. Vis. Comput. Graph. 19 (12) (2013) 2159–2168, https://doi.org/10.1109/TVCG.2013.228.