Visual Analytics of People Movement Data for Abnormal Behavior Pattern Detection

Zhenghao Chen, Jianlong Zhou, Jeremy Swanson, Xiuying Wang, Dagan Feng, Fang Chen

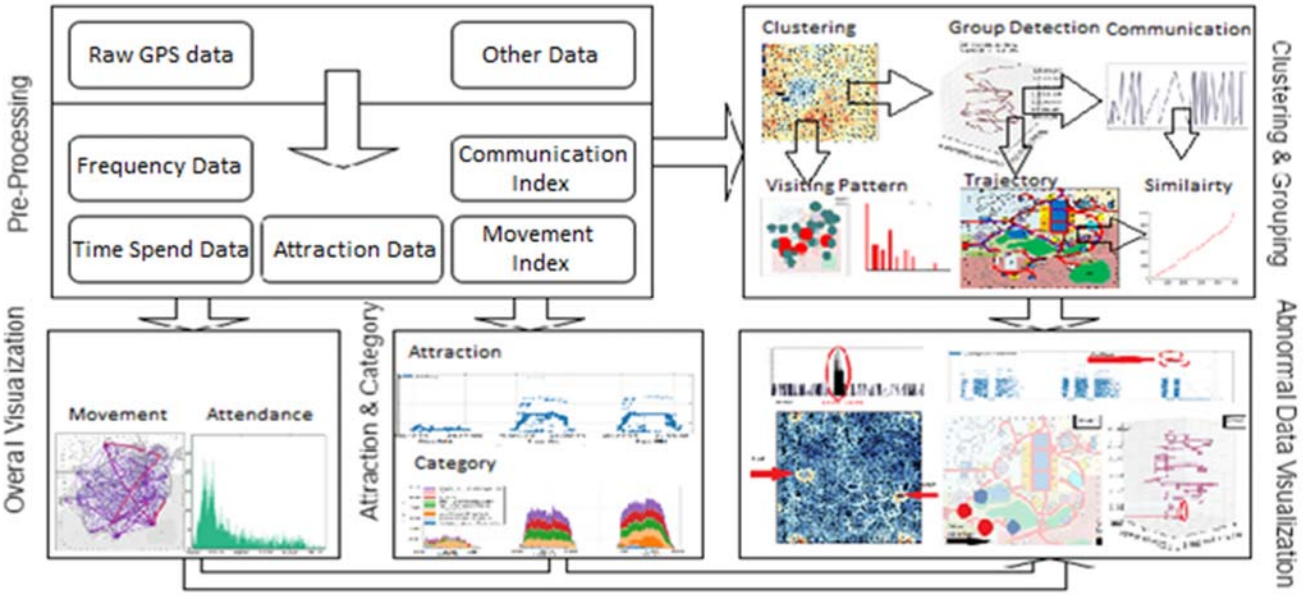


Fig. 1 Overview of the proposed approach: there are five main stages in visual analytics: Pre-processing of raw data, clustering & grouping, overall visualization, attraction visualization, and visual analytics of abnormal behavior pattern.

Abstract— Visual analytics is one of significant approaches to gain insights from movement data. This paper proposes a visual analytics pipeline for trajectory data enabling better understanding movements of visitors, preferences of attractions, similarities of groups, and popularities of attractions. Such understanding helps to extract patterns of abnormal behaviors in the movements. The proposed approach uses Self-Organizing Map (SOM) to cluster people into groups based on different visiting features such as preferences. Further analyses of groups such as similarity of trajectories using Dynamic Time Warping (DTW) are then performed to understand movement patterns. The proposed approach is able to successfully detect abnormal behavior patterns such as crimes in the people data set of a popular amusement park. The proposed approach can be used in the safety related management of events involving a large amount of people movements.

Index Terms— Visual analytics, movement data, Self-Organizing Map, Dynamic Time Warping, crime detection

INTRODUCTION

With the advance of positioning technologies such as sensors and Global Positioning System (GPS), vast amounts of movement data are being generated every second every day from various domains, such as traffic, mobile phones, public transportation, park management, etc. Visual analytics is one of promising techniques widely used to extract knowledge and support better reasoning and understanding of data for planning and decision making in society and business. Taken the people movement data as an example, using visual analytics to understand movement of visitors is crucial for better management of transportation, big events with large number of people, and others involving people movement. Such analytics also helps to address the safety issue and emergency management, specifically avoid occurring of crime, and improve the traffic condition like traffic jam.

Traditional methods for movement data analytics often put too much effort in constructing a sorted database as the back-end and building user interface as the front-end to enable users inspect every single visitor trajectory pattern or aggregate them by certain queries [1]–[6] for visualization. However, these approaches cannot effectively address the analytics of large amount of movement data. Firstly, they only plot the data with various visualizations, but such visualizations cannot be used in an integrated way to help improve the understanding data and get insights from data. Users still need to summarize the visualizations and infer possible causes and results. Secondly, the visualization of single trajectory is not practical especially when the data comes even larger. Recently more advanced approaches such as flow map[7], [8] and trajectory clustering[6] are proposed to analyze movement data. In trajectory clustering, most

- *Zhenghao Chen, Jianlong Zhou, Jeremy Swanson, Fang Chen, CISRO Data61,E-mail:Zhenghao.Chen@nicta.com.au, Jianlong.Zhou@nicta.com.au,Jeremy.Swanson@nicta.com.au, Fang.Chen@nicta.com.au*
- *Xiuying Wang, Dagan Feng University of Sydney, Xiu.Wang@sydney.edu.au, Dagan.Feng@sydney.edu.au*

existing techniques focus on clustering the actual trajectory with classical clustering methodologies[6]. Flow map typically uses arrow symbol to represent the number of objects moving between the places and with clear directions encoding some other information like number of items to represent the movement pattern and trajectories. Both approaches are straightforward and essentially suitable for movement pattern which is relatively simple and regular. However, when trajectory pattern becomes complicated and irregular, and especially movements of huge number of people get massively cluttered, the legibility of flow map would be significantly decreased while clustering such irregular trajectories would also difficult.

In this paper, we propose a visual analytics methodology that combines both people movement data and communication data for understanding movement patterns. Our aim is to mine movement patterns under different conditions and figure out people's motivation of behaviors with information from multimodalities. Various features such as visiting frequency of a site are extracted from movement data. Advanced machine learning method of Self-Organizing Map (SOM) is then utilized to cluster movement patterns. SOM not only clusters movement to meaningful groups, the visualization of SOM results also helps users to detect differences between groups such as unusual patterns of movement. Furthermore, such unusual patterns of movement are verified with the use of another modality of data such as communications between people. In this paper, the people movement data at a crowed park in the VAST 2015 Grand Challenge [1] are analyzed as a case study to demonstrate the effectiveness of our approaches. SOM is used to cluster visitors of the park based on visitor's preferences of attractions in the park. Such clustering also visualizes the majority preferences and motivations of visitors. Followed by visualizing the visitor pattern of each clusters, we can further infer the visitor types such as youth, seniors and so on[9]. Group behavior of park visitors based on the clustering is also analyzed. For instance, real time trajectory and communication pattern [6] are discovered to show how groups travel and communicate. The similarity of those groups is then evaluated with Dynamic Time Warping (DTW). The proposed approach is used to successfully find the unusual pattern such as possible crime in the park data. Specifically, the contributions of this paper include:

- Incorporate both movement data and communication data into the visual analytics pipeline to understand people movement behaviour and test hypotheses.
- Propose a visual analytics methodology by utilizing SOM for clustering and visualizing patterns of movement.
- Set up a pipeline to successfully detect and test unusual patterns in the movement data.

1 Related Work

Visitor Clustering and Grouping: Classical machine learning clustering algorithms of K-means[6], [8], [9] and K-Nearest Neighborhoods[6] are widely used for clustering trajectory data. However, such clustering methods cannot efficiently render a useful result especially for irregular movement patterns such as the people movement in a park. This is mainly because that K needs to be defined in advance for either K-means or K-nearest Neighborhoods and find an optimal K value is a non-trivial work. Therefore such clustering approach for trajectory data analysis cannot show much meaningful results [1]. These clustering methods are also sensitive to noise especially for massive cluttered movement data. Different from previous work, we use unsupervised approach of Self-Organizing Map [10] to cluster people movement in this paper. Furthermore, most of previous work tries to cluster the movement trajectory directly [1], [7],[5]. However, such clustering become very difficult and even impossible when clustering cluttered and irregular trajectories In our work, we try to cluster moving people based on their preferences of locations [9]. Visitors are clustered by the time spent in certain locations and check-in preferences. This kind of clustering method helps to find people who have the high potential to be in the same groups as they have similar or even the same interests. Based on interests and preferences of clusters, we can further infer what kind of visitors they are.

Cluster visualization is also crucial in analytics of movement data[4]. Most existing visual analytics approaches try to model the trajectory information as movement patterns [6], [4]. Since our movement data with irregular and cluttered trajectories make the entire trajectory pattern meaningless, instead of discovering trajectory flow as most of previous approaches [5], we analyze people movement pattern according to their preference clustering. We also use different approaches[5], [9] to visualize the visitor preference pattern. Besides, other data such as communication, time spent on a site, preference categories are also considered as features for clustering.

Grouping movement data is a typical pre-processing step for trajectory data analysis [11]. After grouping instead of investigating single movement, visual analytics approaches analyze movement based on groups. Groups are usually defined as two types: 1) Groups which have the same or similar type of trajectories. Clustering results are usually defined as groups in this case [8], [9]. 2) Only people who visit sites together having exactly the same spatial trajectory can be defined in a group such as family, couple or friends[4], [6]. In order to find these groups, we normally need to compare every individual trajectory data. That is we need to scan entire trajectory database to calculate the difference of one individual trajectory to every other single trajectories. This method is much more accurate than the first type of grouping but is very time-consuming ($O(n^2)$). In our work, we define groups based on clustering results and group visitors within the same clusters after clustering the data, which narrows the search space. This is because that visitors have potential to be in the same group only if they have the same preferences of visiting.

Trajectory Visualization: Two approaches are widely used to visualize trajectories: 2D map-based approach and 3D time-space cubes [4]–[8], [12]. In 2D map-based visualization, *real time trajectory* can be further represented by revealing movement using animations [4], [5], [7], [12]. Particularly, animations are used widely in traffic trajectory analysis [13], [12]. The animation can demonstrate the order of locations that visitors travel as well as other information such as speed changes in real time. From each frame of an animation, we can see how movement occurs during certain intervals. Recently, another approach based on 2D map visualization called spatial flow [7], [8], [14] is recognized as a useful way to visualize the movement pattern. Spatial-time cube[15] is another approach widely used to display massive trajectory data. However, these approaches cannot perfectly present trajectories when trajectories come cluttered. In this paper, both 2D-map based visualization and spatial-cube are used to visualize single group trajectories.

The *similarity of trajectories* is also important in movement visual analytics. Various approaches are used to compute difference between trajectories including Leven-Shtein distance[1], Euclidean distance, and Dynamic Time Warping (DTW) [11], [16]. This paper uses DTW [17], [18] in trajectory comparison.

2 Definition and data description

2.1 Data

This paper uses the data set from the Visual Analytics Science and Technology (VAST) Challenge 2015 (http://vacommunity.org/VAST+Challenge+2015) as a case study to demonstrate the effectiveness of proposed approaches. This data set was on movement of visitors in a modest-size amusement park, DinoFun World, sitting around 215 hectares with 81 different attractions that can be classified by 7 main categories [20]with Rides, Show & Entertainment, Information & Assistant, shopping, Beer Gardens, Restrooms and Food, where Rides can be further distinguished by Thrill Rides Kiddie Rides and Rides for everyone. All attractions are numbered, named and connected by a visitor pathway throughout the park [21]. Each visitor was tracked by a mobile device that records his/her positions in real time and his/her behaviors and communications. For protecting the privacy of users, devices only recorded two behaviors of customers that movement and check-in [1]. This park hosts thousands of visitor every day. Especially, there were 3357, 6411, and 7569 visitors on 6, 7, and 8 June 2014 respectively as an event “Scott Jones Weekend” was held for celebrating the coming of local star Scott Jones. On Sunday, 8 June 2014, a crime happened in this park and rapidly solved by officials [1].Therefore, in this paper, we use this crime event as a specific case study to show how our approaches are used to detect the crime group and prove our hypotheses.

2.2 Pre-Processing

This data consist of two main parts: 25 million individual movement data (Mini-Challenge 1 (MC1)) and 4 million communication data (Mini-Challenge 2 (MC2))[9]. Specifically, movement data is in the format of <timestamp, person id, behavior type, position>, while communication data is in the format of <Timestamp, call from, call to, location>, and pre-processing of data is the first essential step in this study.

In this study, the data are pre-processed in following ways:

- **Check-in Frequency Data**: Each visitor is represented as an 82-dimension vector, and each dimension of a vector represents each attraction. The value of each dimension represents the check-in frequency of this visitor.
- **Time Spend Data**: Similar as check-in frequency data, an 82-dimension vector for each visitor represents 82 attractions and the value of each dimension of the vector is the time that visitor spent (in seconds) during the day.
- **Attraction Data:** Tree data structure is used to store attractions with their position information as branches rooted by categories they belong to.
- **Movement Data Index:** Movement data are indexed in two ways: visitor and attraction. This indexing allows to easily query information when we analyse trajectory, visit pattern and popularity of attractions as well as other features.
- **Communication Data Index:** Each communication of 2 visitors are indexed by the visitors and time which allows us to see the communication pattern of visitors throughout the time.

The Check-in Frequency Data and Time Spend Data are used to represent the visiting pattern of each visitor in one day, which essentially shows the preference of that visitor. Specifically, we analyze how visitors spent their time in all attractions, time spent on each attraction and check-in frequency for each attraction. The data are used to cluster visitors and find abnormal movement behaviors Tree data structure allows to label the categories and attraction where visitors are currently in given the position data of visitors. Movement data index is used to get information for our visitor-oriented and attraction-oriented study. We also investigate frequency of communication between visitors

3 OUR APPROACH

This section presents details our visual analytics approach for people movement. We firstly give an overview of overall movement in the park followed by visitor clustering by their preference using Self-organizing Map [10], [22]. The visualization of U-matrix [10] of SOM demonstrates clusters in the investigated data. The trajectory and communication pattern of groups are then compared using Dynamic Time Warping[17], [18].

3.1 Overall Visualization

Overall visualization of the movement is used to give an overview of the entire movement and visiting patterns of all visitors in the park. This study visualizes the staying time and overall movement to understand the visitor movement behavior.

Visualization of staying time helps to understand how long visitors mostly spent in the same places before they move to next place. Fig. 3 shows the time spent of all visitors on 6 June 2014. From Fig. 3, we can clearly see that visitors mostly spend about 6 to 9 seconds for the same place (left figure in Fig.3) which means that most of time visitors were moving. And visitors normally spent less than 40 minutes for the same place (right figure in Fig. 3), which means if there are certain visitor spent more than 40 minutes in one attraction, we can consider them as outliers. For those outliers, it is essential to figure out reasons they spent so long in one place as those might be some safety issues.

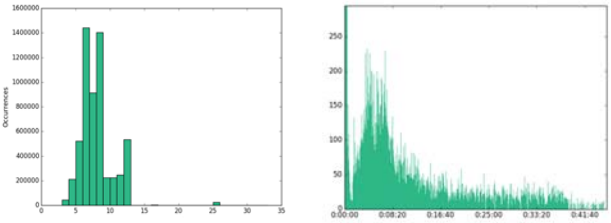


Fig. 2. Time information in same places for all vistors on 6 June 2014: time between points (left), and time spent in same places (right).

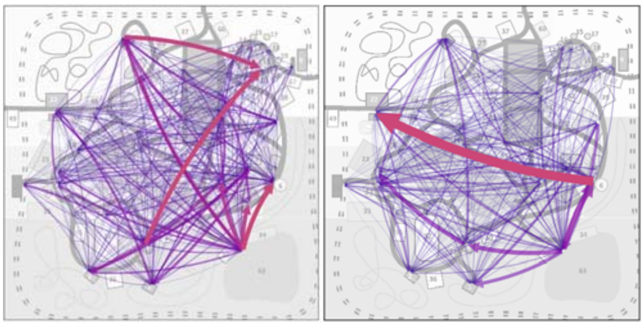


Fig. 3. Visualization of overall movement

Visualization of overall movement provides an overview of visitor movement in the park. Fig. 4 shows the overall movements at different time, where the wider the line width, the more visitor moved between two attractions the line connected. From Fig. 4 (left), we can see that at 11:00 am on 6 June, the two majority groups of movements are from Attraction 11 and 2 to 16, and from Attraction 66 to many other places. At 16:00 pm on 6 June 2014 in Fig. 4 (right), the majority movement is from Attraction 6 to 5 that is the most outstanding movement. Understanding the overall movement of visitors in the park is helpful for the management of the park. For instance, a majority of visitors moved to Attraction 5 at 16:00 pm, more security work needs to be considered at that specific attraction at that time.

3.2 Visiting Pattern Clustering

In order to understand attraction preference of visitors, we cluster visitors based on time they spent and frequency of check- in in different attractions. Essentially, visitors with similar or exactly the same visiting pattern and preference are categorised into the same clusters. From these clusters, we make further hypotheses to understand movement pattern. For instance, if some clusters have the visiting pattern that spent a large amount of time visiting Kiddie Rides and Entertainment Show, we can assume that those clusters can be a big family with kids. Alternatively, if preferences of clusters are mainly distributed in Beer Garden and Thill Rides, those clusters can have high potential to be young friend who are likely to have drink and fun. Based on these observations, visiting clustering can be used to find various movement behaviour patterns. For example, visitors who committed crime might have high probability to have different movement patterns [19] from regular visitors, and therefore visiting clustering can be a helpful to locate them. This paper uses Self-Organizing Map[10] to cluster visiting patterns.

3.2.1 Self-Organizing Map

Self-Organizing Map (SOM) or Self-Organizing Feature Map (SOFM) is an artificial neural network invented by Teuvo Kohonen in 1982 [10]. In the data visualisation area, it is usually used to cluster and visualise multi-dimensional and non-linear data. It includes four main stages which are initialization, competition, cooperation and adaptation.

Initialization: Initialization creates a 2D-array map with nodes in the lattice and each node has an arbitrary value in its different dimensions. Fig. 2 shows how SOM projects a three-dimensional vector onto a 4*4 2D lattice map

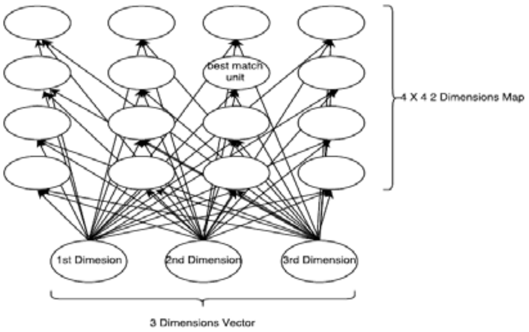


Fig. 4. Initialization of SOM.

Competition: Competition uses Euclidean distance function as in Equation (1) to calculate the weight distance *Dist* between every node $\boldsymbol{W}$ and input vector $\boldsymbol{V}$ in a 2D map. It then defines a node with minimum weight difference as Best-Matching Unit (BMU).

$$Dist = \sqrt{\sum_{i=0}^{i=n} (V_i - W_i)^2} \qquad (1)$$

Cooperation: Cooperation uses radius decay function to locate the neighbourhoods of BMU in each iteration. There are three sub-steps in this stage. The first step is to compute the distance between the node and the BMU with Equation (2).

$$BmuDist = \sqrt{(X_{bmu} - X_{node\ n})^2 + (X_{bmu} - Y_{node\ n})^2} \qquad (2)$$

The second step computes a decay radius of neighborhood circle which takes the BMU as the center as shown in Equation (3). In Equation (3), σ(t) is the decay radius of neighborhood, σ_0 is the initial radius of circle which is equal to the map width, t is the iteration time and λ is a constant number.

$$\sigma(t) = \sigma_0 e^{(-\frac{t}{\lambda})} \qquad t=1, 2, 3 \ldots. \qquad (3)$$

The last step is to compare every node to see whether they are inside their neighborhood circles. If yes or $BmuDist < \sigma(t)$, they are defined as neighbors of related BMU.

Adaptation: Adaptation trains all the nodes inside the neighborhood circle whose center is BMU using neighborhood function to update their weight W as in Equation (4).

$$W(t+1) = W(t) + \Theta(t)L(t)(V(t)\text{-}W(t)) \qquad (4)$$

Where $\Theta(t)$ is used to consider neighborhood in the weight adaptation, L(t) is the learning decay, for computing the learning rate in each training iteration. The factors that affect L(t) are the current iteration time and the initial learning rate L_0. L(t) is calculated with Equation (5).

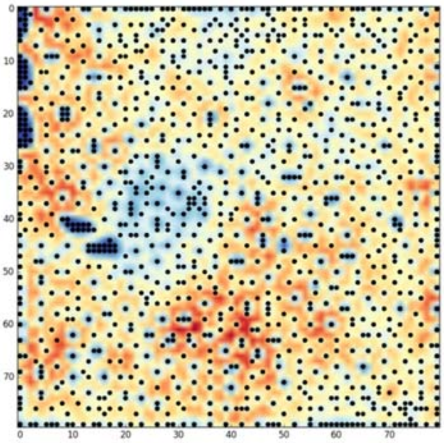
$$L(t) = L_0 e^{(-\frac{t}{\lambda})} \qquad t= 1, 2, 3\ldots.. \qquad (5)$$

The effect of neighborhood on the weight of nodes is the function of the distance between the current node and BMU, the radius of neighborhood circle, and the current training iteration time as shown in Equation (6).

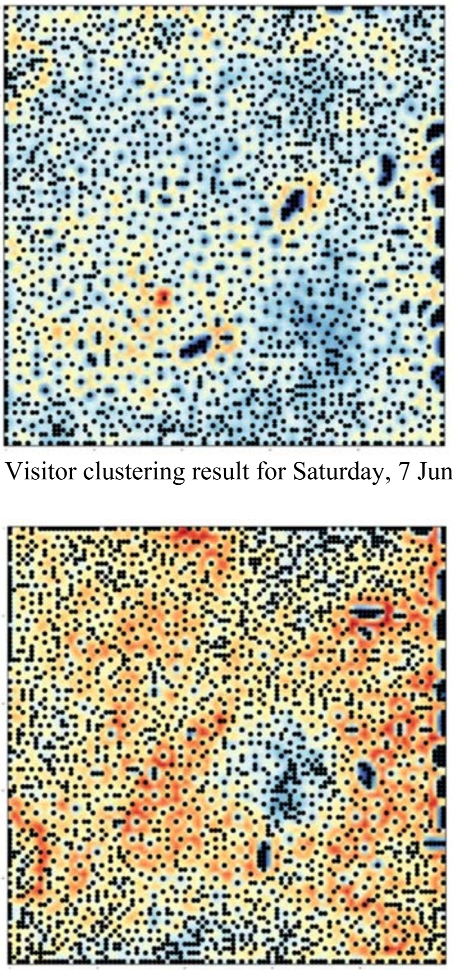
$$\Theta(t) = e^{(-\frac{Dist^2}{2\sigma^2(t)})} \qquad t=1, 2, 3\ldots \qquad (6)$$

After iterative updating of all nodes, the map is self-organized and all nodes save their trained weights. The U-matrix of SOM is used to visualize the cluster map which shows the preference clusters of visitors. In the U-matrix visualization, the outstanding clusters are surrounded by very different and obvious colors which can be easily detected by users.

Fig. 5 shows the U-matrix visualization of SOM based on Check-in Frequency Data and Time Spend Data on Friday (6 June, 2014), Saturday (7 June, 2014) and Sunday (8 June, 2014). Clustering result for Friday used 80*80 2D map for training, while clustering result for Saturday and Sunday used 100*100 2D map for training by considering more visitors of massive data at the weekend than on Friday. From these visualizations, we found that there are 5 large distinct clusters on Friday, while there are 12 and 14 large distinct clusters on Saturday and Sunday respectively.



(a) Visitor clustering result for Friday, 6 June 2014.



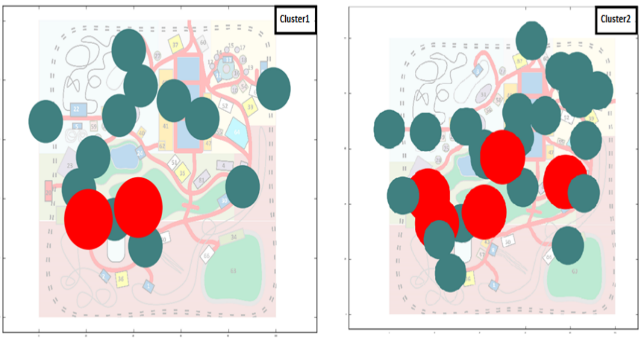
(b) Visitor clustering result for Saturday, 7 June 2014.

(c) Visitor clustering result for Sunday, 8 June 2014.

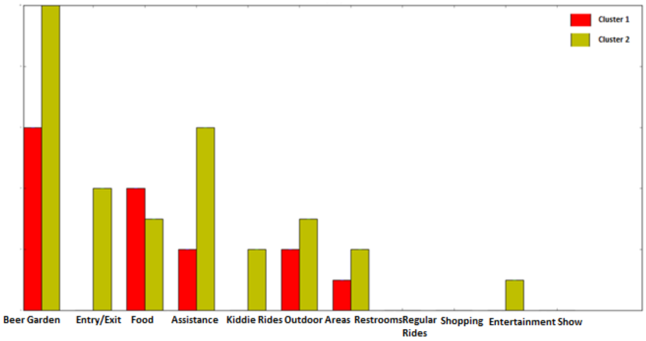
Fig. 5. U-matrix visualization of visitor clustering results for 3 days

3.2.2 Visiting Pattern

This section makes further analyses on clusters in SOM U-matrix to find what kind of visitors are in the same cluster, how long they spent in and how frequent they checked in different attractions and categories to understand clusters.



(a) Visualization of Time Spend and Check-in based on attractions.



(b) Visualization of Time Spend and Check-in based on attraction categories.

Fig. 6: Visiting pattern visualization for two clusters on 6 June 2014.

Fig. 6 shows the visualization of two very distinct clusters on 6 June 2014. As shown in Fig. 6, we can see that Cluster 1 spent most of their time in attraction 65 and 32, while Cluster 2 visited a lot for attractions 53, 65, 32, 35 and 29. In the visualization based on attraction categories as shown in Fig. 6 (b), Cluster 2 clearly spent almost twice in Beer Garden than Cluster 1. They also checked into Show & Entertainment and Kiddies Rides a lot, while Cluster 1 did not even visit these two categories. Based on these observations, we can make an assumption that Cluster 2 is possibly a big family who took kids to Kiddie Rides, watched entertainment shows as well as took family drink. Cluster 2 can potentially be some young friends and they went to the park to drink and have food together. Using this technique, we can visualize all distinct clusters to discover what kind of preferences they had and further hypothesize what type of visitors were in the cluster.

3.3 Group Detection

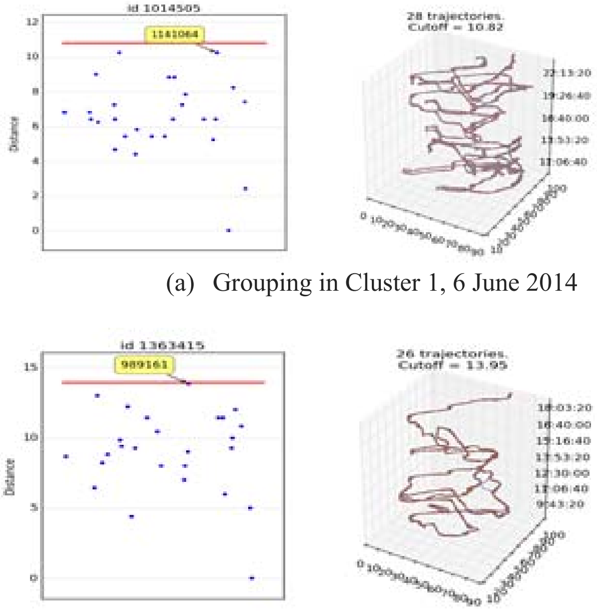
This section further analyses SOM clusters by detecting groups inside a cluster. The analysis is based on the assumption that people visiting a park usually travel with several people or even more together. Such visitors are called a group in this paper. Therefore, such visitors usually have the same movement trajectory and the same or similar visiting pattern. Only visitors in the same clusters can have potential to be in a group, especially for those high distinct clusters in the SOM, which have the high probability to contain a big group. In this group detection stage, we examine whether those distinct clusters are visitors belongs to big visiting groups such as a family. Such visitor grouping in the same cluster is efficient since it do not need to scan whole data set. In this section, group detection in the same cluster is conducted based on three steps: trajectory sampling, communications of visitors, and trajectory comparison with dynamic time warping.

3.3.1 Trajectory Sampling

We compare trajectory of one visitor to trajectories of all other visitors in a cluster. Trajectories firstly need to be resampled to have the same resolution. For example, visitor A may have the movement record slightly different from visitor B. We resample both of them to make sure that in every time step they have the position data$(x_i\,, y_i)$. The Euclidean distance between every time point of two visitors is computed and summed to get the trajectory distance of two visitors as shown in Equation (7).

$$Traj\_Dist = \sum_{i=0}^{i=n} \sqrt{(X_{Ai} - X_{Bi})^2 + (Y_{Ai} - Y_{Bi})^2} \qquad (7)$$

Where Traj_Dist is the trajectory distance between two visitors of A and B, n is the number of time points in a trajectory.



(a) Grouping in Cluster 1, 6 June 2014

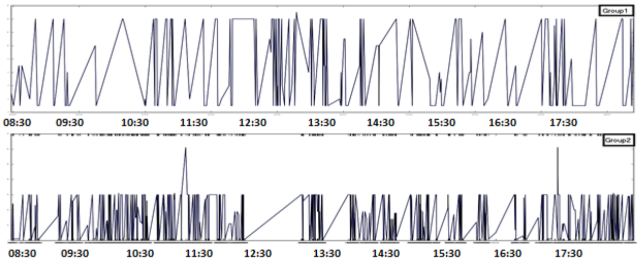
(b) Grouping in Cluster 2, 6 June 2014

Fig.7 Group Detection in Clusters on 6 June 2014

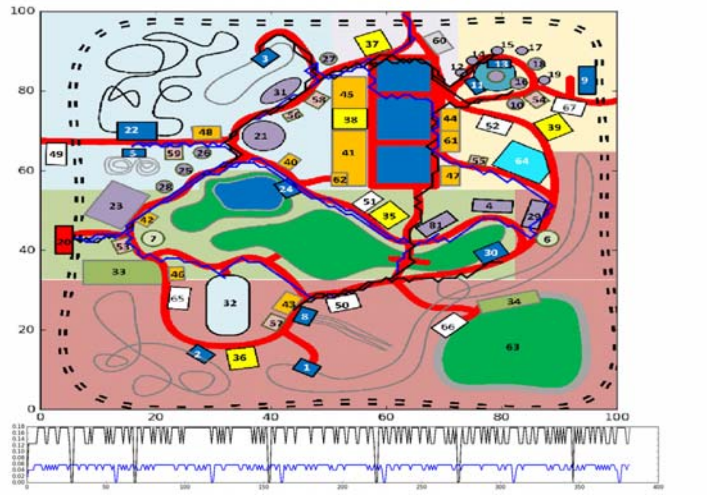
Fig. 7 shows the group detection in two clusters (cluster 1 and cluster 2) we analyzed in Section 4.2.3. In Cluster 1, we can see that the largest distance inside the group is around 11, while the largest distance in Cluster 2 is 12. From 3D spatial-time cube of trajectories [12] as shown in the right of the Fig. 7, we can see that trajectories of visitors in clusters are exactly matched. This grouping detection method enables us to save all visitors in a group using group tree data structure.

3.3.2 Communication and Trajectory

This section visualizes communication information of groups to show relations between visitor communications and their trajectories. Fig. 8 shows the communication frequency in Group 1 and Group 2 (a) as well as the trajectories and speeds of two groups (b). From the figure, we can see that communication inside Group 1 or Group 2 had frequent contacts, while the communication between Group 1 and Group2 turned out to be a straight line with value 0 throughout, which means that there was no communication between Group 1 and Group 2 and these two groups are independent groups.



(a) Communication Pattern of Group 1 and Group 2



(b) Real time trajectory of Group 1 and Group 2

Fig. 8 (a) Communication frequency in group 1 and group 2 on 6 June 2014, (b) Trajectory of two groups (top) and their real time speed (bottom), where the black line represents group 1 with average speed around 1.5m/s and blue line represents group 2 with average speed of 1m/s.

3.3.3 Trajectory Similarity Evaluation

At this stage, two visitor groups have trajectory information and communication information which are both time series data. This section uses Dynamic Time Warping to evaluate the similarities among groups. Dynamic Time Warping (DTW) [16][17] is an algorithm to compute the similarity of two sequence data. It is used to compare the similarity of both trajectories and communication between two groups. Fig. 9 shows the similarity between Group 1 and Group 2 in trajectories and communication.

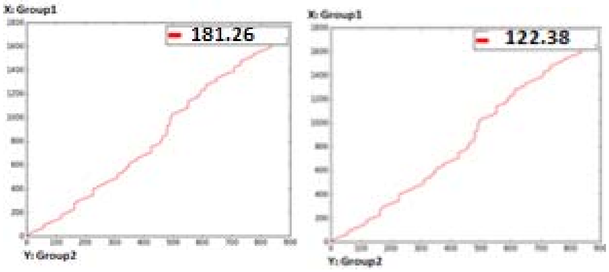


Fig. 9 Similarity of Group 1 and Group 2: (left) similarity of their trajectories, (right) similarity of their communication.

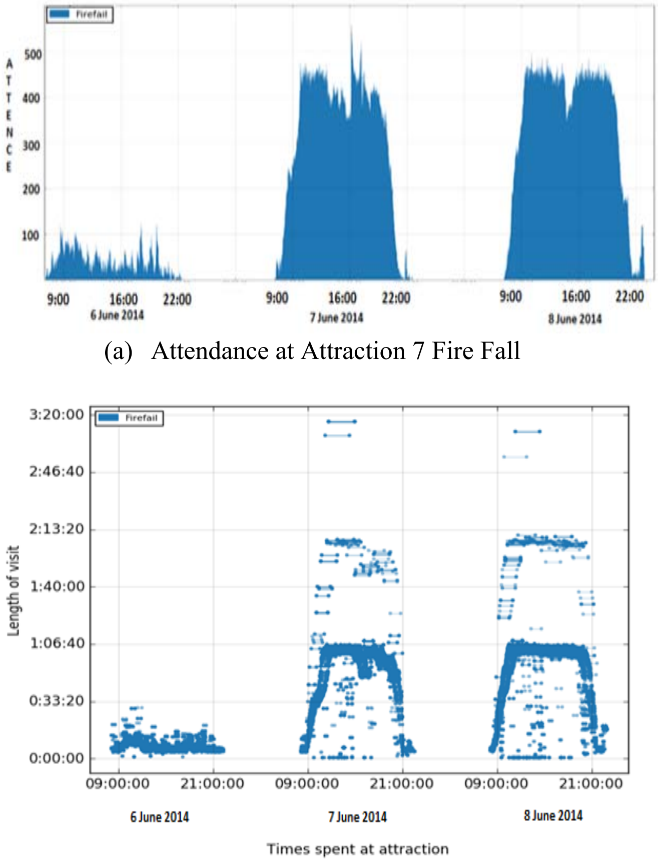
Besides similarities in trajectories and communication, this section also considers other similarities including visitors' visiting pattern, the spent time and frequency of check-in for categories and attractions. This section defines a new similarity distance.

3.4 Attraction and Category Visualization

Attraction-oriented visualization is another important approach to understand how visitors spent time in each attraction and attraction category, two different visualization methodologies are used in this section to show people's visiting preferences.

3.4.1 Attraction Visualization

To study how visitors travel at a certain attraction, we firstly visualize the number of visitors in this attraction throughout the time, and then show how long visitors spent in this attraction. Fig. 10 shows visitor attendance and time spent at Attraction 7 Fire Fall in three days.



(a) Attendance at Attraction 7 Fire Fall

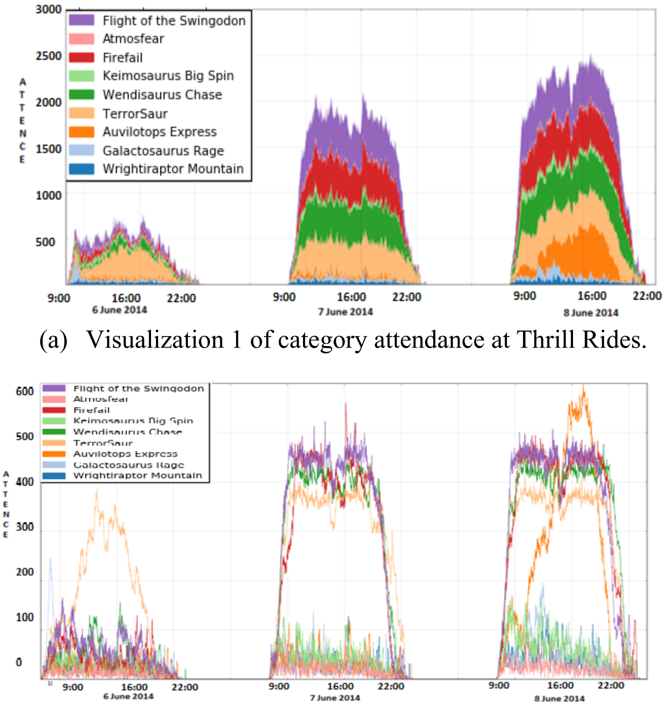
(b) Spend Time at Attraction 7 Fire Fall

Fig. 10 Visiting pattern at Attraction 7, Fire Fall on 3 days that 6 (left), 7 (middle), and 8 (right) June 2014.

From Fig. 10, we can find the distinguished differences in visitor attendance and time spent at Attraction 7 over three days. For instance, the attendance of visitors on Friday (6 June 2014) at this attraction in different was always less than 120 and averaged around 50-60 throughout the day (Fig. 10 (a) left). Also we can see from Fig. 10 (b) left that almost all visitors spent less than 30 minutes at Attraction 7 on Friday throughout the day. However, the weekend (7, 8 June 2014) visiting pattern was much different from Friday, more than 400 visitors attended this attraction most of the time at the weekend and the time they spent at this attraction was mostly half hour to 1 hour. This is reasonable because we can assume that people had more time at the weekend and therefore more visitors came to this attraction and spent longer time. It was also found from the visualization that majority visitors spent less than two hours at the weekend at this attraction. From the spend time visualization in Fig. 10 (b), we also found that there are four outliers at the weekend. These four obvious outliers (visitors) can be related to safety issues, for instance they might be relevant to crime or others in the park or we can even hypothesize they may meet some troubles like getting injured.

3.4.2 Category Visualization

Besides single attraction visualization, we also visualize different attraction categories to understand visiting patterns on different days. These information can reflect which attraction is most popular at different time and what changes of visiting pattern in three days. Fig. 11 shows two visualizations of category attendance at Thrill Rides



(a) Visualization 1 of category attendance at Thrill Rides.

(b) Visualization 2 of category attendance at Thrill Rides.

Fig.11 Visualization of category attendance at Thrill Rides.

From the visualization in Fig. 11, we can see that the attendance pattern of all Thrill Rides over three days are quite different: Attraction 4 Terror Saur was the most popular Thrill Rides on Friday, on 6 June 2014, while at the weekend (7, 8 June 2014) Attraction 81 Flight of the Swingdon seemed to become hottest one with largest visitors.

4 CASE STUDY – CRIME DETECTION

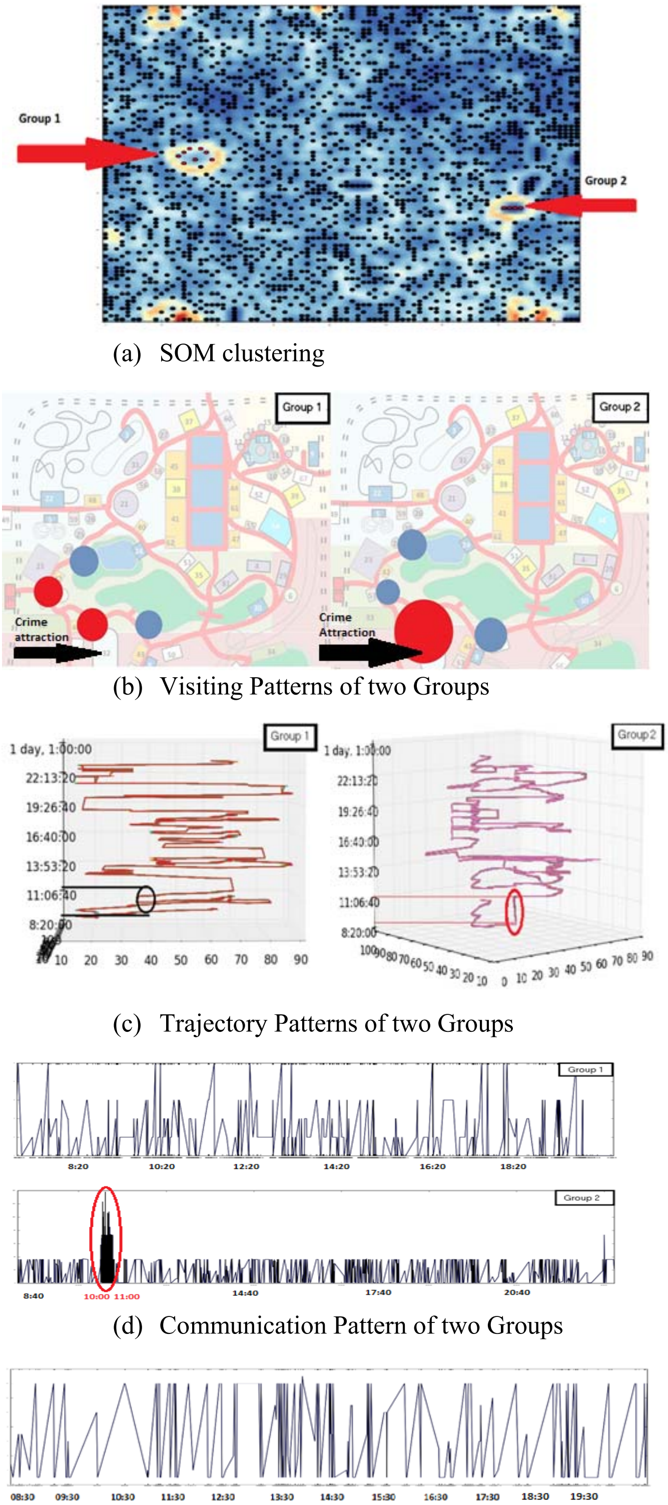
In this section, we present a case study to show how our proposed approaches are used to find the crime group in the DinoFun World. This case study is to detect a crime accident occurred in DinoFun World on 8 June 2014 [20]. By analysing and visualising visiting data of that date using our approaches, we accurately find the crime groups.

4.1 Background and Pre-Processing

In the visiting data of DinoFun World, the local soccer star Scott Jones had a celebration called “Scott Jones Show” at DinoFun World from 6 to 8 June 2014 over the weekend. All his personal honors including an Olympic medal were occurred at Creighton Pavilion (Attraction 32). There was a crime group vandalized the exhibiting, breaking into the place and did crime things on Sunday, 8 June 2014 [21]. The visual analytics approach proposed in this paper is used to detect the crime group.

4.2 Crime Pattern Visualization

The visiting pattern of a crime group was usually different from the regular visitors [19]. Regular visitors usually have random distribution on the time spend and check-in in attractions in the park, while crime person may spend particularly longer time and higher frequency of check-in in certain places which is mostly the attraction that crime occurred. The trajectories of crime groups also may be different from normal visitors, for instance, when the show in Creighton Pavilion is temporarily closed and ready for the next show, normal visitors probably leave the place travelling to another place while crime people more likely still stay there during that time as it is good time to commit crime as no many visitors around there. From these assumptions, our visual analytics firstly conducts the clustering of data in the morning on Sunday. We then visualize the visiting pattern of outstanding clusters, and figure out their trajectory and communication behavior to detect crime groups.



(a) SOM clustering

(b) Visiting Patterns of two Groups

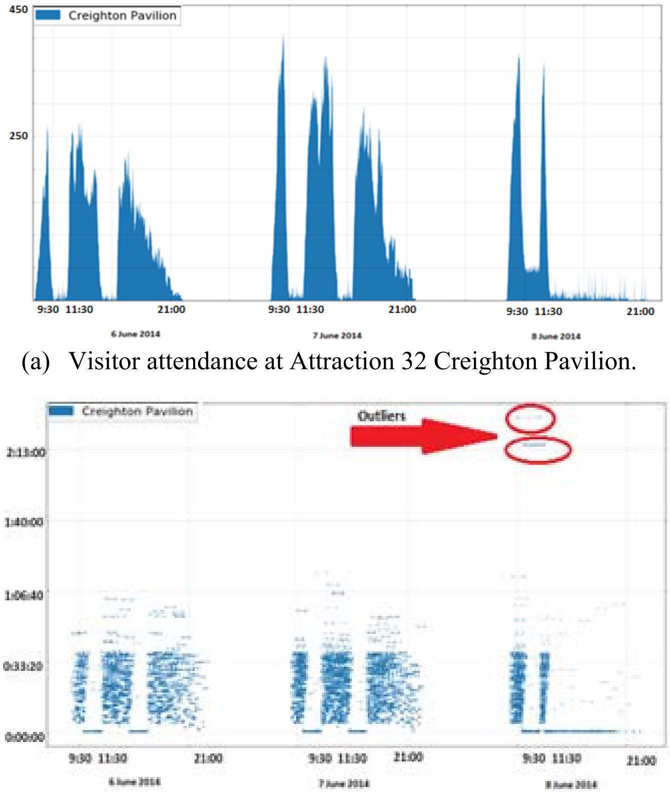
(c) Trajectory Patterns of two Groups

(d) Communication Pattern of two Groups

(e) Communication Patterns Between two Groups

Fig. 12 Crime pattern visualization

Fig. 12 shows the visualization of trajectory and communication visualization of groups in order to detect crime groups. From Fig. 12 we can clearly see crime patterns of groups. In Fig. 12 (a), visualization of SOM Map shows six distinct clusters. By further analyzing these six clusters, we found that among six clusters, two of them have the highest probability to be relevant with each other. We name these clusters as group 1 and groups as shown in Fig. 12 (a). By visualizing visiting patterns of these two groups as shown in Fig. 12 (b) we found that group 1 spent the most of their morning at two attractions of Attraction 53 Smoky Wood BBQ and Attraction 32 Creighton Pavilion, while group 2 spent almost the whole morning at Attraction 32 Creighton Pavilion. As both of groups spent quite a lot of time at the Attraction 32, we suspect these two groups committed a crime in Attraction 32 because of longer time spent at that attraction. By analyzing their movement in spatial-time cube visualization as shown in Fig. 12 (c), we found that from 9:30 am to 11:30 am, group 1 did check-in at Attraction 32 and stayed there until 11:00am, and then moved to Attraction 53. While group 2 spent all the time from about 9:30 am to 11:30 am at Attraction 32 without any further movement. In their communication visualization as shown in Fig. 12 (d), we found that the communication frequency of group 2 was much higher during the period from 10:00 am to 11:00 am than other time period. Therefore, we can assume that the time period from 10:00am to 11:00am was highly possibly the crime time. Based on these observations, we can conclude that group 2 committed crime while we still cannot exclude the suspects of group 1. We further visualize the communication pattern between group 1 and group 2 as shown in Fig. 12 (e), the result shows that group 1 and group 2 kept in touch during the whole day of 8 June 2014, which means that group 1 and group 2 were not independent groups and they knew each other. Therefore, our conclusion is that group 2 committed crime at Attraction 32 while group 1 guarded Attraction 32 and Attraction 53 to help group 2 to commit the crime from about 10:00am to 11:00am.



(a) Visitor attendance at Attraction 32 Creighton Pavilion.

(b) Visitor Time Spend at Attraction 32 Creighton Pavilion.

Fig. 13 Visiting pattern at Attraction 32 Creighton Pavilion.

Visiting data at Attraction Creighton Pavilion is further analysed to confirm our conclusions. Fig. 13 shows the visiting pattern at Attraction 32 Creighton Pavilion. As shown in Fig. 13 (a), we found that there are three large attendance periods on each of three days, which means that "Scott Jones Shows" took place during these three time periods on each day. We also found that from around 9:30am to 11:30am every day, there was a sharp decrease followed by a very small number of check-in of visitors and then increased again. From this pattern, we can infer that from 9:30am to 11:30am this attraction was shortly closed until the next shows ready after 11:30am. Normally during this time period visitors most likely left this attraction. However from the time spend on Attraction 32 as shown in Fig. 13 (b), we found that there were two groups of visitors staying in this place from 9:30am to 11:30am. The time they spent at Creighton Pavilion are clearly outliers as normal visitors spent less than one hour there but they spent more than two hours in this place. Therefore, we can assume that these visitors are related to the crime.

In summary, based on these visual analytics as shown in this section, the whole crime story can be described as follows. "Scott Jones Shows" took place three times every day from Friday to Sunday. After the first show ending at about 9:30am, the attraction was closed until 11:30am. During this time period on Sunday, the crime people were divided into two groups: the first group mainly acted as assistance group and the second group committed the actual crime in Creighton Pavilion. The first group travelled around Attraction 32 to guard the crime. At around 10:00am, the second group started working on breaking into exhibition and stolen. Therefore the communication in the second group was increased a lot. Two groups also kept in touch frequently during the whole crime committing period.

Compared with results from other research [1], our approach found the same results but is more flexible in finding abnormal behavior patterns. The proposed approach in this paper can be implemented in park management for solving safety related issues.

5 VISITING PATTERN PREDICTION

Besides the visual analytics of visiting patterns, this paper also predicts visiting pattern of visitors using their historical data of visiting attractions as features to predict which attraction they are mostly likely about to visit. For example, given a group of visitors having already visited certain attractions, we want to evaluate the probability of all attractions that this group will attend. The prediction models used in this paper include most frequent, Uniform, K-Nearest Neighbours (KNN), Multinomial Naïve Bayes (NB), Decision Tree, Random Forest and Support Vector Machine (SVM). The leave-one-out cross validation is used to evaluate the classification accuracy of prediction models.

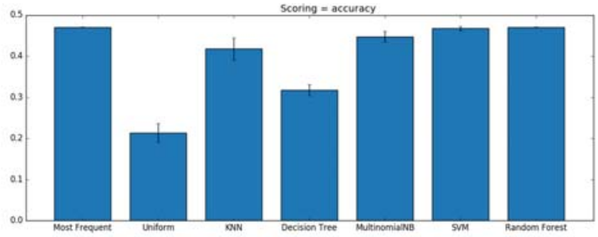


Fig. 14 Result of classification accuracy of different prediction models. we can see that Random Forest outperforms other classifiers in predicting visiting patterns. We got the prediction accuracy at around 48%.

6 CONCLUSION AND FUTURE WORK

The analysis of people movement especially the visitor movement in a park plays significant roles in the park management. This paper proposed using SOM to cluster visitors into groups based on different visiting features such as time spent at attractions and visiting frequencies of attractions. The U-matrix visualization of SOME helped users detect distinct clusters of visitors. Based on these clusters, further visiting information were derived to understand visiting patterns of visitors. The proposed approach was used to detect abnormal behavior patterns such as crime detections. The results showed that the proposed approach can effectively analyze movement data and get insights for human decisions.

Our future work will focus on defining more meaningful features for visiting pattern clustering. Advanced prediction models will also be developed to predict visiting patterns by incorporating both communication information and trajectory data into models.

REFERENCES

[1] M. Steptoe, R. Krueger, Y. Zhang, X. Liang, R. Garcia, S. Kadambi, W. Luo, T. Ertl, and R. Maciejewski, "VAST challenge 2015: Grand challenge - team VADER/VIS Award for Outstanding Comprehensive Submission," in *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2015, pp. 119–120.

[2] B. Yu and B. Zhou, "VAST challenge 2015 solver," in *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2015, pp. 133–134.. Animating images with drawings. In Andrew Glassner, editor, Proceedings of SIGGRAPH '94 (Orlando, Florida, July 24–29, 1994),Com- puter Graphics Proceedings, Annual Conference Series, pages 409–412. ACM SIGGRAPH, ACM Press, July 1994.

[3] A. Puri, D. Liu, S. Chen, S. Fu, T. Wang, Y. Chan, and H. Qu, "ParkVis: A visual analytic system for anomaly detection in DinoFun World," in *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2015, pp. 123–124

[4] G. Andrienko, N. Andrienko, I. Kopanakis, A. Ligtenberg, and S. Wrobel, "Visual Analytics Methods for Movement Data," in *Mobility, Data Mining and Privacy*, F. Giannotti and D. Pedreschi, Eds. Springer Berlin Heidelberg, 2008, pp. 375–410.

[5] N. Andrienko and G. Andrienko, "Visual analytics of movement: An overview of methods, tools and procedures," *Inf. Vis.*, vol. 12, no. 1, pp. 3–24, Jan. 2013

[6] Y. Zheng, "Trajectory Data Mining: An Overview," *ACM Trans Intell Syst Technol*, vol. 6, no. 3, pp. 29:1–29:41, May 2015.

[7] T. von Landesberger, F. Brodkorb, P. Roskosch, N. Andrienko, G. Andrienko, and A. Kerren, "MobilityGraphs: Visual Analysis of Mass Mobility Dynamics via Spatio-Temporal Graphs and Clustering," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 1, pp. 11–20, Jan. 2016.

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

[9] J. Zhao, G. Wang, J. Chae, H. Xu, S. Chen, W. Hatton, S. Towers, M. B. Gorantla, B. Ahlbrand, J. Zhang, A. Malik, S. Ko, and D. S. Ebert, "ParkAnalyzer: Characterizing the movement patterns of visitors VAST 2015 Mini-Challenge 1," in *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2015, pp. 179–180.

[10] T. Kohonen, "The self-organizing map," *Proc. IEEE*, vol. 78, no. 9, pp. 1464–1480, Sep. 1990.

[11] L. Chen, "Similarity Search over Time Series and Trajectory Data," University of Waterloo, Waterloo, Ont., Canada, Canada, 2005.

[12] M. Lu, C. Lai, Y. Tangzhi, J. Liang, and X. Yuan, "Visual analysis of route choice behaviour based on GPS trajectories," in *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2015, pp. 203–204.

[13] X. Huang, Y. Zhao, J. Yang, C. Zhang, C. Ma, and X. Ye, "TrajGraph: A Graph-Based Visual Analytics Approach to Studying Urban Network Centralities Using Taxi Trajectory Data," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 1, pp. 160–169, Jan. 2016.

[14] P. Bak, H. J. Ship, A. Yaeli, Y. Nardi, E. Packer, G. Saadoun, J. Bnayahu, and L. Peterfreund, “Visual analytics for movement behavior in traffic and transportation,” *IBM J. Res. Dev.*, vol. 59, no. 2/3, pp. 10:1–10:12, Mar. 2015.

[15] T. Kapler and W. Wright, “GeoTime information visualization,” in *IEEE Symposium on Information Visualization, 2004. INFOVIS 2004*, 2004, pp. 25–32.

[16] “Haase, Jens, and Ulf Brefeld. ‘Finding similar movements in positional data streams.’ Proc. ECML/PKDD Workshop on Machine Learning and Data Mining for Sports Analytics. 2013.”

[17] “Dynamic Time Warping,” in *Information Retrieval for Music and Motion*, Springer Berlin Heidelberg, 2007, pp. 69–84.

[18] S. Salvador and P. Chan, “Toward Accurate Dynamic Time Warping in Linear Time and Space,” *Intell Data Anal*, vol. 11, no. 5, pp. 561–580, Oct. 2007.

[19] S. V. Nath, “Crime Pattern Detection Using Data Mining,” in 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops, 2006. WI-IAT 2006 Workshops, 2006, pp. 41–44.

[20] “VAST Challenge 2015.” [Online]. Available: http://vacommunity.org/VAST+Challenge+2015. [Accessed: 01-Mar-2016].

[21] M. Whiting, K. Cook, G. Grinstein, J. Fallon, K. Liggett, D. Staheli, and J. Crouser, “VAST Challenge 2015: Mayhem at dinofun world,” in *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2015, pp. 113–118.

[22] V. Moosavi, “Computing With Contextual Numbers,” *ArXiv14080889 Cs*, Aug. 2014.