Trajectory Pattern Identification
and Anomaly Detection of Pedestrian Flows
Based on Visual Clustering

Li Li$^{(\boxtimes)}$ and Christopher Leckie

Department of Computing and Information Systems, The University of Melbourne,
Parkville, Melbourne, VIC 3010, Australia
`lli10@student.unimelb.edu.au`, `caleckie@unimelb.edu.au`

Abstract. Extracting pedestrian movement patterns and determining anomalous regions/time periods is a major challenge in data mining of massive trajectory datasets. In this paper, we apply contour map and visual clustering algorithms to visually identify and analyse areas/time periods with anomalous distributions of pedestrian flows. Contour maps are adopted as the visualization method of the origin-destination flow matrix to describe the distribution of pedestrian movement in terms of entry/exit areas. By transforming the origin-destination flow matrix into a dissimilarity matrix, the iVAT visual clustering algorithm is applied to visually cluster the most popular and related areas. A novel method based on the iVAT algorithm is proposed to detect normal/abnormal time periods with similar/anomalous pedestrian flow patterns. Synthetic and large, real-life datasets are used to validate the effectiveness of our proposed algorithms.

Keywords: Data mining · Pedestrian trajectory pattern · Visualization · Clustering · iVAT algorithm

1 Introduction

There is growing interest in the problem of extracting useful information from massive trajectory datasets derived by various sensing methods. Understanding patterns of pedestrian movement is useful in applications such as pedestrian flow management, public security and safety. A major challenge in pattern analysis of pedestrian movement is how to discover and describe the movement patterns hidden in trajectories, and identify any misbehaviour or interesting events.

The main approaches to trajectory data analysis and anomaly detection fall into the category of trajectory data mining. To detect and recognize social events, two common approaches used to address this problem are statistical methods combined with classification and clustering-based methods. Many existing approaches to address this problem have the limitations that they focus on the

details of trajectories, but do not consider the characteristics of the trajectory distribution.

In this paper, we address this limitation of existing approaches by proposing the use of contour maps and visual clustering. Contour maps are a very useful visualization tool for three-dimensional data, which we adopt to visually describe the connection between different subareas and describe the distribution of trajectories. Visual clustering methods such as VAT (Visual Assessment of cluster Tendency) and iVAT (improved Visual Assessment of cluster Tendency) [1,9] are proposed to visually assess the clustering tendency of a set of objects. By using the VAT/iVAT approach, we are able to visualize and determine the possible number of clusters of locations or the periods with similar activity, and then determine abnormal areas/days with significantly different trajectory distributions.

The main contributions of our paper are as follows. First, we use a visualization method to describe pedestrian movement distributions in terms of their origin and destination points. By transforming the origin-destination flow matrix into a dissimilarity matrix, we visually cluster the most popular flows using the VAT/iVAT algorithms. The popular flow patterns are important for monitoring the safety and security of public areas. Second, we propose a novel method of detecting abnormal time periods with anomalous pedestrian trajectory distributions based on the results of the VAT/iVAT algorithms. By doing so, it is possible to detect the occurrence and impact of special events. Finally, we evaluate our methods and make relevant comparisons on a large, real-life dataset, the Edinburgh informatics forum database [7], and demonstrate the effectiveness of our proposed algorithms.

2   Related Work

An important aspect of a monitoring system is to detect significant events or unusual behaviour in that environment. By mining pedestrian trajectories, it is possible to detect the occurrence of major events, such as celebrations, parades, business promotions, accidents, disasters and others, which may be threats to public security or safety. With the increasing availability of big trajectory data, there have been various research methods on the detection and recognition of anomalous social events from trajectory data. In this paper, we focus on the challenges of how to detect normal patterns of flows, and how to detect abnormal pedestrian flow patterns.

The first problem we address is to find and visualize the most visited subareas in a given region. In [3], an algorithm for extracting popular regions was proposed, with the popular regions defined as regions with trajectory densities larger than a given threshold, which is manually prescribed instead of adaptively adjusted. In this paper, we aim to find a parameter free algorithm for this problem. Liu et al. [4] designed real time analytical method for spatial-temporal data of daily travel patterns in metropolitan urban environments. The data for analysis of Liu [4] are taxi traces and smart card records, whereas in contrast

we mainly focus on pedestrian trajectories. While they provided an analysis on travel patterns, they did not consider the problem of detecting anomalies. Liu et al. [5] and Lu et al. [6] shared the same idea of extracting entry/exit points and using length analysis to derive indoor scene structure and identify abnormal motion behaviours. However, the methods in [5,6] only detect particular instances of trajectories. In this paper, using the same dataset [5,6], we focus on using the origin-destination matrix as a whole to describe the pedestrian flow patterns and cluster the most visited areas.

The second problem we address is how to detect a set of time periods with abnormal trajectory motion patterns. In the literature, there have been many research studies on anomaly detection methods for traffic or pedestrian data. Pang et al. [8] proposed a statistical model, which adapts likelihood ratio tests to find anomalous regions for monitoring the emergence of unexpected behaviour based on GPS data from taxis. Chawla et al. [2] adopted Principal Component Analysis (PCA) to detect traffic anomalies from GPS data. In [2], moving activities of a crowd were simulated as the movements of a group of points, and the distribution of point groups is described with fractal dimensions. PCA was used to remove the disturbed factors from a feature vector and maintain only relevant information. Witayangkurn et al. [10] proposed a framework based on a hidden Markov model to construct a pattern of spatial-temporal movement of people in each area in a grid during each time period. However, they focused on using changes in the population to detect anomalies. In this paper, we intend to detect anomalous time periods of trajectory distributions based on the use of visual clustering methods, which can partition the data into clusters in a visual manner. In this way, we can provide a robust, unsupervised approach for clustering periods of normal pedestrian activity and visually highlighting periods of anomalous pedestrian activity.

3   Problem Statement

Our aim is to find the most popular routes of pedestrians in a given region and check which areas are the most related by visualizing the motion patterns, and to detect and visualize abnormal time periods by comparing the distributions of motion patterns. An outlier or anomaly in a dataset is considered to be an inconsistent observation (or subset of observations) compared with the remainder of that set of data, such as a substantial change in the popularity of pedestrian routes.

Suppose that the structure of the monitored area is known and can be divided into k subareas $A = \{A_1, A_2, , A_k\}$ according to the functions of different areas or some other predefined labelling. For example, as shown in Fig. 1 (from the Edinburgh informatics forum database), there are 13 functional areas in this forum.

The dataset consists of a set of detected people walking through these subareas. Suppose we are given the trajectory data $T = \{T_{day_1}, T_{day_2}, ..., T_{day_m}\}$ covering a set of m days $D = \{day_1, day_2, .., day_m\}$, where on day_i a set of trajectories
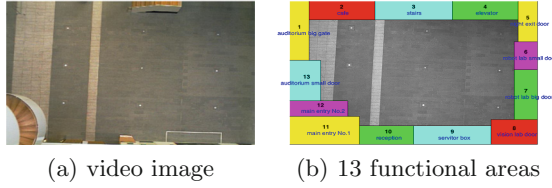
(a) video image (b) 13 functional areas

Fig. 1. Video image and functional areas of the Edinburgh informatics forum

$T_{day_i} = \{T_{i1}, T_{i2}, ..., T_{in_i}\}$ has been collected, where n_i is the number of trajectories on day_i, and it is assumed that the trajectory data is the outcome of a robust tracking system. Each trajectory is composed of triplets (x, y, t) containing X/Y coordinates and sampling time $T_{ij} = \{(x_1, y_1, t_1), (x_2, y_2, t_2), .., (x_l, y_l, t_l)\}_{ij}$, where l is the number of points in trajectory T_{ij}. Given trajectory dataset T, our aim is to find and visualize the most visited subareas $A' \subset A$ and identify a set of time periods $D' \subset D$ with abnormal trajectory motion patterns.

To address this aim, we examine three related research questions: (1) how to identify and summarize related sets of pedestrian flows over a given time period; (2) how to identify which time periods exhibit similar patterns of pedestrian flows; and (3) how to identify which time periods have experienced anomalous flow patterns? We present our approach to these research questions in the following three sections.

4 Case Study - Edinburgh Pedestrian Flow

The trajectory dataset from the University of Edinburgh is used as a case study in this paper. The dataset provides the scene under surveillance and the configuration of it and the image covers most of the main hall, which has been shown in Fig. 1. The most significant features of the hall are that there are many entry and exit points, i.e., the main entrance to the building, lifts, access to the Atrium, access to the second part of the hall, staircase, reception desk, and the four other exits, which means that there are a variety of possible pedestrian flows in this area.

The dataset consists of a set of detected targets of people walking through the Informatics Forum, the main building of the School of Informatics at the University of Edinburgh. The valid data covers 118 days of observation, resulting in about 90,000 observed trajectories in total. Substantial differences are observed between weekdays (Mon-Fri) and weekends. The average number of trajectories on weekdays is 932, which is significantly larger than the number of 140 on weekends.

Some papers [5–7] provided clustering methods to exact the pedestrian flows from trajectory data in terms of these entry and exit points. Based on the pedestrian trajectories detected from video images in Majecka et al. [7], which are represented by a sequence of centroid positions, we provide a visual clustering using contour maps and iVAT.

5   Summarizing Related Flows

The first question we address is how to identify and summarize related sets of pedestrian flows over a given time period. The challenges in addressing this question are how to summarize a given pedestrian flow matrix so that a user can identify the dominant flows, and then how to identify related subsets or clusters of flows. In this way, we can summarize the pedestrian activity over a given period of time. In this section, we illustrate methods to summarize a flow matrix and detect normal/abnormal days by identifying related flows.

5.1   Synthetic Cases

The rows and columns of a contour map reflect the order in which entry/exit pairs are labelled, but it does not reflect the clustering relationships between flows, so we would like to visually group related flows. Consider the following synthetic example.
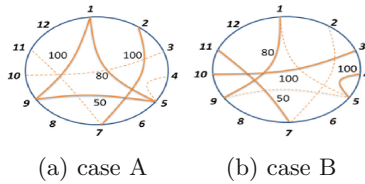

(a) case A   (b) case B

Fig. 2. Two synthetic cases. Italic numbers from 1 to 12 represent 12 different areas. Lines and dashed lines between different areas represent the large pedestrian flows and small pedestrian flows respectively, and numbers next to lines indicate the size of pedestrian flows (Color figure online)

As shown in Fig. 2, these two synthetic cases both have four relatively high pedestrian flows 50, 80, 100 and 100, but the relationships between flows are different. To visually display the origin-destination pair distribution characteristics, we introduce contours to represent the flow matrix. Contours indicate equal valued regions with the same colour. This is similar to a heat map, and the characteristics of the matrix distribution can be visually analysed. The application of contours helps us better visualize and compare distribution characteristics of the origin-destination flow matrix, which would otherwise be hard to analyse only by the values of the matrix. The contour maps of the two cases are shown in Fig. 3 respectively.

Although we can easily find the distribution of trajectories flows and detect origin/destination pairs that have high pedestrian flows in Fig. 3, we cannot easily identify related flows, since the light areas are scattered on the contour map. The main challenge is how to reorder the rows/columns of a contour map to group related flows. In this paper, we propose to treat this as a visual clustering problem.
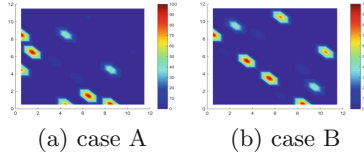
(a) case A   (b) case B

Fig. 3. Contour maps of the two synthetic cases in Fig. 2

VAT and iVAT are useful tools for visual assessment of clustering tendency, as is shown by Bezdek et al. [1] and Wang et al. [9]. The VAT algorithm displays a reordered dissimilarity matrix D as a grey-scale image with a modified version of Prims minimal spanning tree algorithm. The iVAT algorithm augments VAT by applying a path-based distance transform to the input dissimilarity data before VAT images are made. It reorders the dissimilarity matrix of the given set of objects so that it can display any clusters as dark blocks along the diagonal of the image, and a diagonal dark block appears in the iVAT image only when a tight group exists in the data. In this paper, we only provide the results based on the iVAT algorithm. The main steps of iVAT are:

Step 1: Transform input dissimilarity matrix $D \rightarrow D'$ using a path-based distance;

Step 2: VAT is applied to reorder $D' \rightarrow D'^{*}$, resulting in an iVAT image $I(D'^{*})$ whose $(i,j)^{th}$ element is a scaled dissimilarity value between objects o_i and o_j.

Since a dissimilarity matrix D is the input data to the iVAT algorithm, a method to transform the origin-destination matrix F to a dissimilarity matrix D is proposed in our paper. Considering that the origin-destination flow matrix F is non-symmetric (F_{ij} and F_{ji} may be different), the first step is to transform the flow matrix to be symmetric. There are three methods to derive a symmetric matrix S: (1) $S_{ij} = S_{ji} = max(F_{ij}, F_{ji})$; (2) $S_{ij} = S_{ji} = min(F_{ij}, F_{ji})$; (3) $S_{ij} = S_{ji} = (F_{ij} + F_{ji})/2$.

This symmetric flow matrix S can be normalized by using $S'_{ij} = S_{ij}/S_{max}$, where S_{max} is the value of the largest element in S. Then we can compute the dissimilarity matrix D. If $i \neq j$, $D_{ij} = 1 - S'_{ij}$; otherwise, $D_{ij} = S'_{ij}$, where S' is the normalized symmetric transferring matrix. When $i = j$, the dissimilarity between the same area is 0; when $i \neq j$, the dissimilarity between two areas decreases as the normalized symmetric flow matrix S' increases, which means that high pedestrian flows result in low dissimilarity values, and vice versa.

To verify the effectiveness of iVAT, we apply it to the two synthetic examples. The iVAT image results and reordered contour maps are in Fig. 4. The reordering of these two cases are both Areas {1 9 5 2 7 11 3 10 4 6 8 12}. However, the clustering results are different. For case 1, the clustering is {(1 9 5) (2 7) 11 3 10 4 6 8 12}, i.e., Areas 1, 9 and 5 are strongly related and also 2 and 7. For case 2, the clustering is {(1 9) 5 2 (7 11) (3 10) (4 6) 8 12}. The results indicate that iVAT can cluster the related areas correctly.
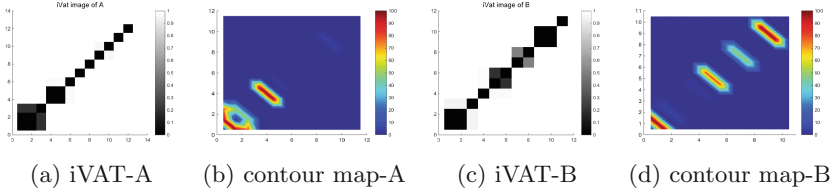
(a) iVAT-A   (b) contour map-A   (c) iVAT-B   (d) contour map-B

Fig. 4. Results on synthetic examples

Table 1. Origin-destination flow matrix

Area	1	2	3	4	5	6	7	8	9	10	11	12	13
13	0	8	0	0	0	0	0	2	0	0	0	0	0
12	1	17	0	0	0	0	0	0	0	0	3	0	0
11	4	102	5	2	0	0	0	11	0	11	0	4	1
10	0	3	1	0	3	0	0	0	0	0	11	1	1
9	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	16	29	0	6	0	2	0	0	0	14	0	3
7	0	0	0	0	1	0	0	1	0	0	0	0	0
6	0	0	0	0	1	0	0	0	0	0	2	0	0
5	0	7	26	17	0	0	0	4	0	1	0	0	0
4	0	3	8	0	13	0	0	0	0	0	1	0	0
3	0	1	0	8	25	0	0	16	0	2	20	0	0
2	0	0	2	5	5	0	0	32	0	4	103	21	2
1	0	2	0	0	0	0	0	0	0	0	0	0	0

Table 2. Reordered origin-destination matrix

Area	1	11	2	8	3	5	12	4	10	13	6	7	9
9	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	1	0	1	0	0	0	0	0	0	0
6	0	2	0	0	0	1	0	0	0	0	0	0	0
13	0	0	8	2	0	0	0	0	0	0	0	0	0
10	0	11	3	0	1	3	1	0	0	1	0	0	0
4	0	1	3	0	8	13	0	0	0	0	0	0	0
12	1	3	17	0	0	0	0	0	0	0	0	0	0
5	0	0	7	4	26	0	0	17	1	0	0	0	0
3	0	20	1	16	0	25	0	8	2	0	0	0	0
8	0	14	16	0	29	6	0	0	0	3	0	2	0
2	0	103	0	32	2	5	21	5	4	2	0	0	0
11	4	0	102	11	5	0	4	2	11	1	0	0	0
1	0	2	0	0	0	0	0	0	0	0	0	0	0

5.2   Case of Real Trajectory Data

Next, we test our method of visual clustering on the real trajectories from one Sunday (20-Jun-2010). We assume that the structure of the scene is known, so we can classify trajectories based on the location of their first (start) and last (end) regions. For example, given $T = \{(x_{start}, y_{start}, t_{start}), ..., (x_{end}, y_{end}, t_{end})\}$ with $(x_{start}, y_{start}, t_{start}) \in A_{11}$ and $(x_{end}, y_{end}, t_{end}) \in A_6$, then T belongs to (11,6).

By counting all the trajectories, we obtain an origin-destination flow matrix (i.e., the frequency-adjacency matrix) F. In this matrix, the value of $F(i,j)$ represents the number of trajectories which start at A_i and end at A_j. Note that the origin-destination matrix can be asymmetric, e.g., Table 1 on Sunday (20-Jun-2010). Applying the iVAT algorithm, there are 12 clusters for all 13 areas, as is shown in Fig. 5. The clustering result is {1 (11 2) 8 3 5 12 4 10 13 6 7 9}.

Using the iVAT results, we obtain the reordered origin-destination flow matrix, which is shown in Table 2. For most origin-destination pairs, there are few trajectories between them (S_{ij} is small compared with S_{max}), leading to $S'_{ij} \sim 0$ and $D_{ij} \sim 1$. Thus, most clusters/area groups contain only one area, except the cluster containing A_{11} and A_2, which means Areas A_2 and A_{11} have high pedestrian flows and they are most related to each other.

6   Identifying Time Periods with Similar Flows

The second question we address in this paper is how to identify which time periods exhibit similar patterns of pedestrian flows. The challenges in addressing this question are how to compare the flow patterns of different time periods, and how to identify which time periods have similar flow patterns. This enables users to profile normal activity.

6.1   Comparing Flow Patterns

Given flow matrices F_i from time period T_i and F_j from time period T_j, we require a measure of how similar are the flows between these two time periods, i.e., we require a distance measure $d(F_i, F_j)$. We use the Frobenius norm which reflects the pairwise difference of individual flows between the same pairs of location, $\|F_i - F_j\|_F = \sqrt{Tr\left[(F_i - F_j)(F_i - F_j)^T\right]}$, where $(F_i - F_j)^T$ is the transpose of $F_i - F_j$, and Tr means the trace of the matrix. For example, given $F_i = \begin{pmatrix} 0 & 5 \\ 3 & 2 \end{pmatrix}$ and $F_j = \begin{pmatrix} 2 & 1 \\ 2 & 4 \end{pmatrix}$, then the Frobenius norm $\|F_i - F_j\|_F = \sqrt{Tr\left[\begin{pmatrix} 20 & -10 \\ -10 & 5 \end{pmatrix}\right]}$.

6.2   Identifying Similar Time Periods

Given a set of flow matrices $F = \{F_1, F_2, ..., F_m\}$ corresponding to m different time periods $T_1, T_2, ..., T_m$, we would like to group or cluster these flow matrices so that we can identify which time periods have similar flow patterns. For example, if F contains seven flow matrices, each corresponding to the average flows on each day of the week, then we would like to detect F in order to identify which days have similar pedestrian traffic.

To achieve this goal, we again make use of the iVAT algorithm. First, we create a $m \times m$ distance matrix D_F, where the $(i, j)^{th}$ entry in D_F is $d(F_i, F_j) = \|F_i - F_j\|_F$. The distance matrix can be normalized by using $norm(D_F) = (D_F - min(D_F))/(max(D_F) - min(D_F))$, where $min(D_F)$ and $max(D_F)$ are the minimum and maximum value in D_F respectively. We then reorder the normalized D_F using iVAT to produce D'_F, which should visually reorder the clusters of time periods with similar flow patterns.

We evaluated our proposed method on the data set of 118 days, which has been classified into seven groups, corresponding to 7 days of the week. Then the averaged flow matrices of the 7 days of the week (7 samples, i.e., Sun, Mon, Tue, Wed, Thu, Fri, Sat) are compared. The iVAT results are shown in Fig. 6. The ordering of the iVAT image is $\{(4\ 6\ 3\ 5\ 2)\ (1\ 7)\}$, and it shows two clusters, corresponding to clusters of weekdays (Wed, Fri, Tue, Thu, Mon) and the weekend (Sun, Sat).
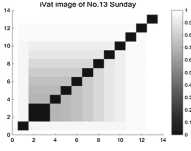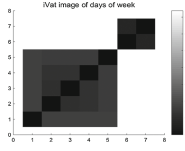
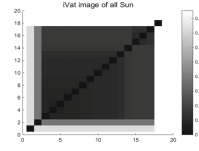**Fig. 5.** iVAT of No.13 Sun   **Fig. 6.** iVAT of 7 days   **Fig. 7.** iVAT of all Sun

7   Identifying Anomalous Flow Patterns

Once we have a profile of normal flow patterns over different periods of time, our final question is how to identify which time periods have experienced unusual or anomalous flow patterns. The challenge in addressing this question is how to identify individual time periods in which the pedestrian flows significantly differ from what is expected. This enables users to detect when an anomaly has occurred, and to analyse how the pedestrian flows during that time period differ from what is expected.

Given a set of flow matrices $F = \{F_1, F_2, ..., F_m\}$, the aim of visual anomaly detection is to detect a subset of these flow matrices that are anomalous or outliers compared to the rest. As before, we use the Frobenius norm to compare flow patterns from different time periods, and construct D_F. We then reorder the distance matrix D_F using iVAT to generate D'_F. When we visualize D_F, any anomalous time periods should appear as singleton dark blocks, which are significantly different form the larger clusters in F. For example, consider the set of flow matrices for all Sundays, the iVAT result is shown in Fig. 7. The iVAT result shows that Sunday has four clusters $\{(13)\ (4)\ (8\ 15\ 16\ 5\ 9\ 17\ 18\ 3\ 12\ 11\ 7\ 10\ 1\ 6\ 2)\ (14)\}$, which means that the anomalous time periods are $D' = \{Sun_{13}, Sun_4, Sun_{14}\}$. The average of normal Sundays contour map, and No.4, No.13 and No.14 Sunday contour maps are shown in Fig. 8.
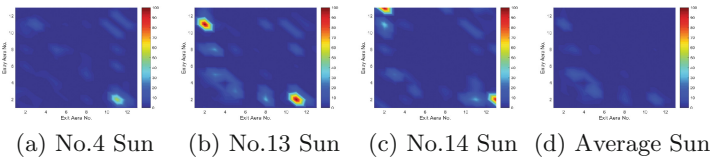

(a) No.4 Sun   (b) No.13 Sun   (c) No.14 Sun   (d) Average Sun

Fig. 8. Contour of abnormal and normal Sundays (Color figure online)

The contour maps indicate that each of the three anomalous Sundays has different high value regions, and all these three time periods are significantly different from the distribution of the average of normal Sundays. For example, in Fig. 8(b), the top left area and lower right area are very bright, which indicates that there are lots of people moving between A_2 and A_{11} as the most visited

areas, corresponding to the two largest values $F(11,2) = 102$ and $F(2,11) = 103$ in the origin-destination flow matrix respectively. Some other areas are rather dark, indicating few people moving between these area pairs. There are some regions with relatively bright colour, indicating relatively high value of trajectory numbers between corresponding area pairs, e.g. $A_2 \rightarrow A_8$, $A_5 \rightarrow A_3$, $A_8 \rightarrow A_3$. Also, similar analysis can be applied to other days of week, and the results are omitted to save space.

8   Conclusions and Future Work

We have used the origin-destination matrix to discover and characterize the connectivity between places or regions. In order to find and visualize related areas, we introduce a contour map to represent the origin-destination flow matrix, and propose a visual and parameter-free area clustering method based on the VAT/iVAT algorithms. To detect and visualize abnormal days with significantly different flow patterns, an iVAT based method is also developed. The results on synthetic data and the Edinburgh informatics forum database show that our methods can effectively cluster related areas and identify normal/abnormal pedestrian flow patterns. Possible future research directions are to discuss on scalability of the method on large data and to modify the proposed method for data stream analysis.

References

1. Bezdek, J.C., Hathaway, R.J.: VAT: a tool for visual assessment of (cluster) tendency. In: IJCNN 2002, pp. 2225–2230 (2002)
2. Chawla, S., Yu, Z., Jiafeng, H.: Inferring the root cause in road traffic anomalies. In: IEEE ICDM 2012, pp. 141–150 (2012)
3. Giannotti, F., Mirco, N., Fabio, P., Dino, P.: Trajectory pattern mining. In: KDD 2007, pp. 330–339 (2007)
4. Liu, L., Assaf, B., Carlo, R.: Urban mobility landscape: real time monitoring of urban mobility patterns. In: Proceedings of the 11th International Conference on Computers in Urban Planning and Urban Management, pp. 1–16 (2009)
5. Liu, W., Xinyi, C., Pengfei, H., Norman, I.B.: Learning motion patterns in unstructured scene based on latent structural information. J. Vis. Lang. Comput. **25**(1), 43–53 (2014)
6. Lu, X., Caixia, W., Nader, K., Arie, C., Anthony, S.: Deriving implicit indoor scene structure with path analysis. In: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness, pp. 43–50 (2011)
7. Majecka, B.: Statistical models of pedestrian behaviour in the forum. Master's thesis, School of Informatics, University of Edinburgh (2009)
8. Pang, L.X., Sanjay, C., Wei, L., Yu, Z.: On detection of emerging anomalous traffic patterns using GPS data. Data Knowl. Eng. **87**, 357–373 (2013)

9. Wang, L., Uyen, T.N., James, C.B., Christopher, A.L., Kotagiri, R.: iVAT and aVAT: enhanced visual analysis for cluster tendency assessment. In: PAKDD 2010, pp. 16–27 (2010)
10. Witayangkurn, A., Teerayut, H., Yoshihide, S., Ryosuke, S.: Anomalous event detection on large-scale GPS data from mobile phones using hidden Markov model and cloud platform. In: UbiComp. 2013, pp. 1219–1228 (2013)