Identifying Pickpocket Suspects from Large-Scale Public Transit Records

Nidamanuri Srinu[1], Voora Yogitha Lakshmi[2], Neelisetty Rajyalakshmi[3], Jampu Sandhya[4], Battula Manoj Kumar[5], Indla Sarath Naga Sai Venkatesh[6]

1 Assistant professor, Dept Of CSE, Qis College of Engineering and Technology, Ongole, Prakasam (Dt)
2, 3, 4, 5, 6 Students, Dept Of CSE, Qis College of Engineering and Technology, Ongole, Prakasam (Dt)

Abstract: Data mining is useful in a wide variety of applications, the most unique and popular applications such as anomaly/suspect detection from the large surveillance system. The abnormality detection is performed on various datasets, which has unique challenges and issues. Outlier detection from the Global Positioning System (GPS) and dynamic trajectory data is much more complicated due to its dynamic update nature. This paper reviews the related works of various suspect, anomaly and outlier detection schemes on mobility datasets. Several works proposed in the literature to detect the abnormality from the traveling behavior which is collected from various wireless media. The unique features among high dimensional dynamic datasets are always developed different types of issues. Detecting and analyzing such dynamic features for finding anomaly is carried out by many types of research in different ways such as cluster oriented, context-aware, dimensionality reduction techniques, graph approaches. After a successful analysis of these approaches, this survey gives an outline to tackle the problems of those techniques.

Keywords Automated Fare Collection; Travel Behaviors; Mobility Patterns; Public Safety; Anomaly Detection.

I. INTRODUCTION

Public transit passengers can easily become distracted in crowded environments, where they are often rushing from one location to another. Having their focus drift from their belongings, they often become common targets of pickpockets [1, 2]. During the first 9 months of 2014, it was reported that 350 pickpockets were apprehended in the subway system and 490 on buses in Beijing.1 Many other big cities around the world, such as Barcelona, Rome, and Paris, also suffer from pickpocket problems.2 Indeed, it is challenging to detect theft activities committed by cunning thieves who know how to escape without being discovered. It is critical to provide a smart surveillance and tracking tool for transit system security personnel.

With rapid advances in information technology and infrastructure, transactional records collected by automated fare collection (AFC) systems are now available for understanding passengers' mobility patterns and urban dynamics [3, 4, 5, 6, 7]. Most existing studies focus on identifying regular, collective mobility patterns, such as commute flows and transit networks. Our study is the first to focus on identifying thieves based on AFC data. It is possible to detect thieves using AFC records because behavioral differences logged in the mobility footprints may be used to separate suspects from regular passengers. Examples of such behaviors include traveling for an extended length of time, making unnecessary transfers, and taking regular routes with random stops. Designing an intelligent system that automatically extracts specific, identified behavioral features and dynamically detects and tracks pickpocket suspects has become a possibility.

Fig. 1: Example trajectories of passengers.

Detecting thieves based on AFC records is not a simple outlier detection problem. Fig. 1 shows the difference between a known thief and an outlier. We can see a number of trajectories between hot regions A and B. By careful examination, we see that most passengers move from one region to another using a near-optimal configuration (e.g., shortest time/distance, or a minimal number of transfers). However, a passenger (a known suspect) who took the path A -> C-> D->! B looks suspicious because there is no need to make transfers at C and D in order to reach B. Based on the above observation, passengers who exhibit such abnormal behaviors will be selected for further examination. In contrast, another passenger who travels from E to B is an outlier, since few passengers take the same path. However, this passenger is likely just a regular

passenger who originates from a less crowded area. Detecting thieves is challenging also because not every trip made by a regular passenger looks normal. Regular commuters may occasionally make trips to visit friends or places of interest, and such trips may look suspicious by how much they deviate from regular passenger behaviors.

Adding to this complex landscape, a large number of AFC records are being collected from millions of passengers, when only a tiny fraction of passengers are actual pickpockets. Pinpointing such a small group of people within such a large scale dataset is analogous to searching for a needle in the haystack. Meanwhile, we need to effectively transform our knowledge based on model development into a decision support system [8]. Such a system needs to provide real-time decision recommendations to guide security personnel to perform their work more efficiently.

To this end, in this paper, a complete method is taken to meet the above demanding situations. Especially, we first construct a function representation for profiling passengers. Furthermore, we set up a two-step framework to split normal movement patterns from abnormal behaviors, and eventually, distinguish thieves from ordinary passengers. Ultimately, we leverage real-international datasets from more than one source for version education and validation, and put into effect a prototype device for give up customers.

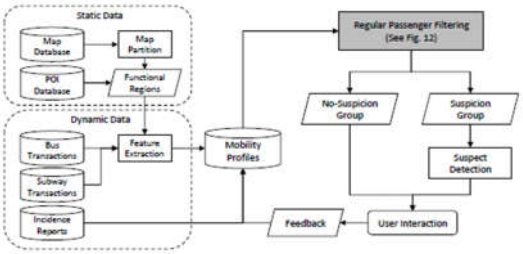


Fig. 2: The overall framework.

Figure 2 indicates the general architecture of our framework. We first partition the city area into areas with functional classes. Then, the mobility characteristics of passengers are extracted from transit information and incident reports. Furthermore, we construct a man or woman mobility database to keep the profile of each passenger. Subsequent, we implement our framework by using everyday passenger filtering and suspect detection. The system is green and interactive, with both mobile and laptop clients. Sooner or later, the user feedback statistics, which include newly showed thieves, will be entered as floor reality for future model education.

II. RELATED WORKS

As urban sensing records, inclusive of GPS lines, call detail records, and smart card logs, develop ubiquitous, studies efforts committed to reading such records have led to a quantity of works in latest years. in the context of mining public transportation statistics, in this segment, we provide a quick assessment of the associated paintings.

A. Passengers Activity Patterns

The primary organization of present literature specializes in finding patterns in passenger activity facts. Such knowledge may be beneficial in an expansion of packages, and performs a critical position in efficiently locating and fulfilling passenger desires. Examples encompass assessing the performance of the transit community, figuring out and optimizing intricate or incorrect bus routes, enhancing the accuracy of passenger float forecasted between two areas, and making carrier adjustments that accommodate variations in ridership on exclusive days. Especially, [4] envisioned the crowdedness of diverse stations within the transportation community using AFC information. [9] Measured the variability of transit behaviors on one of kind days of the week. similarly, special research have investigated precise traits of touring patterns of the aged [10], college students, and adults [9], which furnished interesting insights for know-how behavioral differences of sub-populations.

It's been advised that human mobility patterns observe a high diploma of spatial and temporal regularity, and are for this reason fantastically predictable [11, 12]. Through figuring out trip styles, those studies normally aimed to discovering motion styles by locating frequently visited locations of everyday passengers, who traveled the same sequence of places at a similar time of day. For example, [13], [14] identified spatiotemporal patterns from GPS lines of taxis for night bus path making plans.

Current research that hit upon anomalies in urban sensing records may be divided into categories: the ones based on places, and people on trajectories. Along the line of vicinity-based anomaly detection, [15] presented a framework that discovered the context of various useful areas in a metropolis, which supplied the premise of our feature extraction approach (see phase 3.2). Similarly, [16] attempted to find out informal relationships among spatiotemporal outliers. [17] Mined consultant terms from social

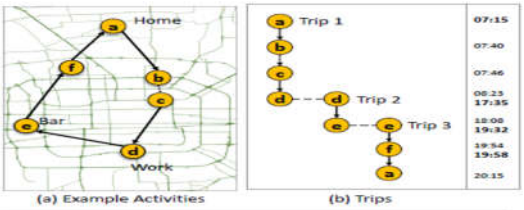
media posts when location-applicable activities happened inside the city, consisting of accidents or protests. [18] Located black-hole or volcano patterns in human mobility facts in a town, which can speedy, identify collecting events, such as football fits or concert events. Detection of such anomalies can help send alerts, and offer enters for intelligent choice aid, including smoothing the visitors float [18]. The main purpose of trajectory primarily based anomaly detection is to discover a small percent of individuals, whose motion lines are uniquely specific from the overall populace. One example is to perceive fraudulent taxi riding behaviors. A massive number of research have investigated trajectory based anomaly detection using facts mining techniques, together with graph primarily based [3], clustering based totally [4, 5, 19], neighborhood/context-aware based totally [20, 21], measurement reduction primarily based [22], and proof based totally (e.g., the usage of Dempster-Shafer principle [23]). At the same time as the trajectory of pickpockets containing features which are implicit, previously unknown, and probably beneficial from huge datasets, pickpocket suspect detection based totally on AFC data is a unique problem that has now not been considered inside the literature, and proves to be a hard research enterprise.

III DATA DESCRIPTION

A. Transit Records

In this study is based on a large-scale transit records dataset collected from a public transit system that includes buses and subways. Passengers utilizing the transportation service are calculated charge by the distance they travel. An enables secure access smart card is issued to each passenger, who has to swipe the card when they board or exit a vehicle. The AFC methods then measure the fare according to the stops of boarding and exiting. As a result, each raw AFC record consists of the smart card ID, the route number, the event (i.e., boarding or exiting), the station, and the time stamp. We transformed the information so that each transportation record consists of one boarding and one exiting event of the same ID. After deleting replicates and tremendously infrequent riders, we are left with over 1.6 billion records that involve nearly 6 million passengers.

With the purpose of this approach describe the data and consequent feature extraction process clearly; here we simplify two concepts, transportation records and trips, using a real example. A) Is the original trajectory on the city map; Part (b) Separates the trajectory into three split trips; and Part (c) describes the corresponding transportation records in our dataset.



(a) Example Activities (b) Trips

Smart Card ID	Route Number	Boarding Station	Boarding Time	Exiting Station	Exiting Time
4322	Route 52	a	07:15	b	07:40
4322	Route 26	c	07:46	d	08:23
4322	Route 11	d	17:35	e	18:08
4322	Route 11	e	19:32	f	19:54
4322	Route 16	f	19:58	a	20:15

(c) Transit Records

Fig. 3: An example of trips and transit records.

Intuitively, a transit file corresponds to one segment of a passenger's transit among a pair of consecutive boarding and exiting occasions. Even though this segment of transit may skip some of stations, the passenger does now not go out the car throughout this time. In comparison, a trip consists of one or more such segments, which connect places of hobby (i.e., wherein the passenger stays for prolonged intervals of time) on the 2 ends.

Definition 1 (Transit report). A transit document tr incorporates the following statistics:

- tr_{route}: the bus/subway direction wide variety,
- tr_{sboard}, tr_{tboard}: the boarding station and time,
- tr_{sexit}, tr_{texit}: the exiting station and time.

As an end result, for each transit file, we have been capable of compute the travel distance tr_{dist}, tour time tr_{time}, and wide variety of stops tr_{stops} all through the transit.

Definition 2 (journey). A trip Tr is a chain of transit facts $T_r = (tr^1, tr^2,\ldots, tr^n)$, where the passenger's foundation vicinity is $Tr_{origin} = tr^1$vboard and the vacation spot is $Tr_{dest} = tr^n{}_{sexit}$.

In practice, we assemble one ride report if and handiest if the time gap among consecutive transit statistics is 30 minis or less. The journey's time length is calculated as $Tr_{time} = tr^n{}_{texit} - tr^1{}_{tboard}$.

B. Geographical Information

To project the transit statistics to the metropolis map, we made use of external datasets of vital geographical statistics, namely, the road network information, the factors of hobbies (POI) statistics, and the general public transit stations. The POI statistics consist of the geo-coordinates of organizations and landmarks with their categories.

We bear in mind ten huge classes of POI, as summarized in table 1.

TABLE 1: Categories of POI and frequencies.

Category	Examples	Frequency
Home	Apartment buildings	28,731
Work	Office buildings	71,364
Education	Schools, training centers	3,527
Food	Restaurants and dining	56,906
Shopping	Shopping malls and outlets	24,310
Entertainment	Museums, theaters, clubs	18,223
Scenic Spot	Parks, sports fields	2,362
Transportation	Airports, transit centers	15,287
Healthcare	Hospitals, pharmacy	8,685
Car services	Car sales, repairs	1,781

As a preprocessing step, we first accompanied Yuan's work [15] to segment the city location into small regions with the aid of most important street networks, as proven in Fig. 4(a). Then, using the POI statistics, we categorized each vicinity into one of the ten practical zones we identified in desk 1. These regions are then color coded visualized in Fig. 4(b).

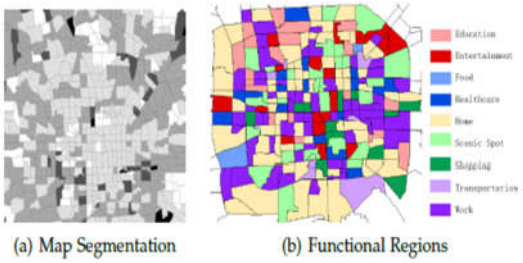


Fig. 4: Geographical information.

The general public transit network dataset presents the geo-co-ordinates of the diverse bus and subway stations at the road networks. In overall, we've got 44,524 bus stations (points in gray) covered via 896 bus routes, and 320 subway stations (points in blue) blanketed via 18 subway routes. To remove the redundancy and better model the mobility styles, we merged stations located at the same avenue intersection.

IV. MOBILITY CHARACTERISTICS

To differentiate pickpocket suspects from ordinary passengers, we extracted some of capabilities from passengers' AFC information. In this segment, we describe those capabilities and talk their potential use for characterizing travel patterns in the public transit system.

A. travel Time and Frequency

The each day tour time is described as the full period spent by way of every passenger in the public transit gadget. The everyday driving frequency is described as the range of transit information traveled with the aid of each passenger consistent with day. Indeed, pick pocketing is hard work: a thief has to spend a long term in crowded buses, subways, or stations to perceive smooth targets. To discover more robbery possibilities, a pickpocket tends to live in the public transit system for a long time and makes random, common transfers.

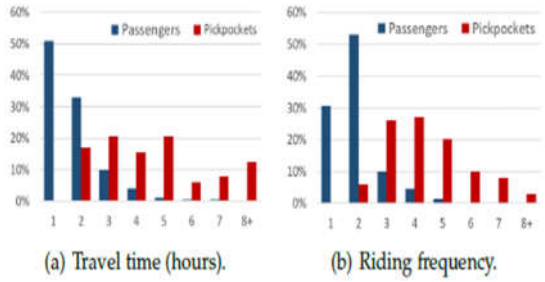


Fig. 5: Distributions of travel time and frequency.

Fig. 5 plots the distribution of every day journey time and riding frequency, respectively. We will see that extra than eighty% of passengers finish their travels inside 2 hours and inside 2 transit facts in step with day. In assessment, the recognized thieves often spend more than three hours touring each day, with a better daily using frequency.

B. brief-Distance Rides

A brief-distance ride is considered a transit phase that passes less than 3 station stops. Because of the density of stations in Beijing, ordinary passengers usually take rides that bypass a bigger range of stations. In comparison, pickpockets regularly switch routes within a few station stops to keep away from attracting fellow passengers' interest and being recognized.

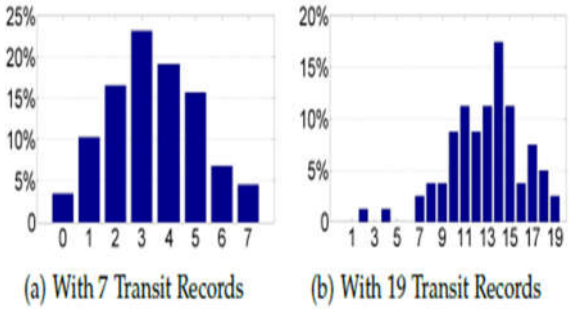


Fig. 6: Distributions of passengers with short-distance trips.

Fig.6 shows the distribution of passengers taking various numbers of short-distance rides, for those with 7 and 19 transit segments, respectively. In each plot, the x-axis is the range of quick-distance rides and the y-axis is the percentage of passengers. For passengers with 7 transit segments, as shown in Fig. 7, the distribution is approximately Gaussian with

suggest three. For people with 19 transit segments, the peak of the distribution is shifted to the right, showing that the relative frequency of short distance rides increases.

C. Detecting Suspects in Real Time

For real-world implementation, we are able to straightforwardly estimate the 2-step version offline each day. With data newly available normal, the re-anticipated fashions can provide higher identity overall performance. However, this kind of naive update system isn't always efficient enough in large-scale facts sets. To make certain that the system is realistic for real-world usage, we adopted a real-time implementation of the regular passenger filtering step (see the shaded container in Fig. 2) with a dynamic ensemble mechanism, as illustrated in Fig. 7.

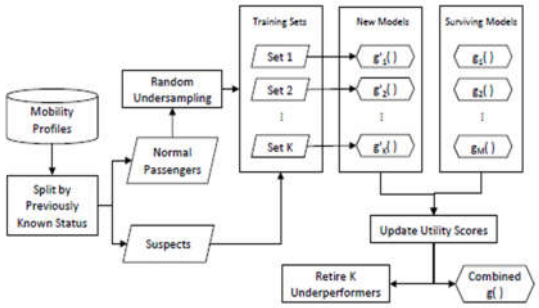


Fig. 7: The incremental update process for training g(.), the regular passenger filtering classifier.

more mainly, we improve the efficiency of normal passenger filtering with the aid of keeping a huge wide variety of base filtering models gm(.), for m = 1, 2,..,M. instead of computing those base filtering models independently every day, we dynamically replace the ensemble to enhance the efficiency. With the software described, we dynamically rank the cutting-edge base filtering fashions and replace the worst candidates with models envisioned with newly to be had information. Our implementation replaces models with the lowest software rankings.

V. EXPERIMENTAL RESULTS

In this phase, we gift experimental outcomes employing our proposed framework. First, we describe the experimental environments and provide implementation info. We then demonstrate the effectiveness of our framework via comparing it to several baseline methods.

A. Experiment Settings

In this subsection, we are able to define our experimental environments and take a look at design. This consists of a brief description on the platform, the baseline techniques, and the performance metrics.

TABLE 2: A performance comparison.

	Offline				Online				
Algorithm	Precision	Recall	F-score	Run Time(s)	Precision	Recall	F-score	Init. Time (s)	Update Time (s)
Classification Methods									
DT	0.002	0.451	0.004	44.81	0.017	1.000	0.034	165.86	15.52
LR	0.003	0.476	0.006	36.72	0.067	0.931	0.124	1643.46	56.17
SVM	0.005	0.512	0.009	21.31	0.049	0.931	0.093	2384.15	62.34
Anomaly Detection Methods									
LOF	0.004	0.560	0.009	300.01+	0.037	0.746	0.071	9821.25+	516.54+
OCSVM	0.015	0.583	0.029	39.67	0.099	0.931	0.179	1845.68	45.56
Two-Step Methods									
LOF+DT	0.011	0.780	0.022	301.18+	0.087	0.795	0.157	9821.25+	428.87+
LOF+LR	0.016	0.829	0.031	301.16+	0.093	0.871	0.168	9821.25+	358.79+
LOF+SVM	0.027	0.758	0.052	318.16+	0.097	0.926	0.175	9821.25+	483.57+
OCSVM+DT	0.053	0.878	0.099	41.19	0.065	0.754	0.120	987.23	63.28
SVM+LR	0.059	0.855	0.110	85.43	0.047	0.931	0.090	2384.15	112.12
OCSVM+SVM	0.071	0.927	0.133	41.05	0.097	0.891	0.175	987.23	65.74
LR+SVM	**0.117**	**0.931**	**0.207**	65.72	**0.169**	**0.931**	**0.288**	**1643.46**	74.35
LR+DT	0.093	0.985	0.170	45.87	0.120	1.000	0.214	1643.46	71.21
DT+SVM	0.086	0.925	0.157	37.54	0.114	0.931	0.203	1274.69	64.23
SMOTE/SVM+LR	0.043	0.845	0.082	71.69	0.041	0.845	0.078	1984.65	112.12
SMOTE/LR+SVM	0.103	0.931	0.185	60.57	0.135	0.754	0.228	1453.54	74.35
SMOTE/LR+DT	0.074	0.952	0.137	43.87	0.094	0.931	0.171	987.32	71.21
SMOTE/DT+SVM	0.082	0.913	0.150	35.45	0.085	0.891	0.155	985.54	64.23

Platform. All offline experiments have been carried out on a windows Server 2012 64-bit gadget (4-CPU, every with 2.6GHz with Quad-middle, and 128G predominant memory). The real-time device changed into implemented on a Spark cluster with 10 nodes. Every node has a Intel i7-4790 CPU3.6GHz CPU with eight cores, 2*8GB Kingston reminiscence, 2 TB SATA3.0 difficult drive, with a Centos 6.five Operation gadget. All algorithms and our actual-international gadget had been developed with Java and Scala.

information instruction. All experiments have been conducted on actual-world datasets described in phase three. There had been approximately 1.7 billion statistics amassed among April and June in 2014. We eliminated passengers from the training set whose most quantity of every day records is no more than 3. After getting rid of duplicates and extraordinarily rare riders, we had over 1.6 billion statistics final that contain about 6 million passengers over the three-month length.

For comparing the actual-time device, we permit the fashions train incrementally and reuse statistics over the years. Every day, there are approximately 14 million records accrued from round five million people.

Baselines. Our approach is as compared with an expansion of competing strategies grouped into the following categories:

- Type techniques. The classification methods, consisting of logistic regression (LR), choice timber (DT) [23], and SVM [22], are straightforwardly match to the education set and evaluated with the take a look at set. Given that the percentage of fantastic times is extraordinarily low, the type hassle is unbalanced, and we expected to look at high type II blunders.

Anomaly Detection. Anomaly detection strategies, such as one-elegance SVM (OC-SVM) [24] and local outlier component (LOF) [20], seem extra appropriate for our problem. Amongst them, LOF is unsupervised, finding outliers via measuring the nearby deviation of given information point with respect to its friends. OC-SVM may be geared up in a supervised way, with simplest the bad instances inside the education set, to perceive the suspects.

- Two-Step strategies. As formerly cited, our technique is a TS approach, together with a poor sample filtering after which applying a conventional type step. For particular aggregate of techniques, we experimented some of opportunities.

Evaluation Metrics. Precision, take into account, and the F-score have been computed primarily based on the take a look at set to evaluate the performances of various techniques. Precision is the number of successfully identified positives divided with the aid of the range of diagnosed positives times. Take into account is the wide variety of efficiently identified positives divided by the variety of all high-quality instances in the check set. And ultimately, the F-rating is calculated as:

$$\text{F-score} = 2 \times \frac{precision \times recall}{precision + recall}.$$

B. Modeling Performance

Table 2 summarizes the performances of our approach (with diverse two-step settings) and the baselines in each offline and actual-time structures. Inside the offline mode, we arrived at numerous interesting observations from the modeling performance. First, the precision of 1-step methods is commonly low. In comparison, all the -step combos significantly improve the precision, with the OCSVM-SVM setup exhibiting the satisfactory performance. This statement suggests that the two-step approach can efficaciously reduce the fake-positives. Second, -step methods additionally finished better in phrases of take into account and the F-score. Finally, the precision of all methods are low. This changed into intentional given that we wanted to make certain an excessive recall. In spite of everything, our floor reality (i.e., flagged suspects) simplest is composed of these confirmed pickpockets, whereas it is possibly there are numerous more pickpockets that go unapprehended.

C. Parameter Tuning

One of the maximum essential parameters we needed to decide was the quantity of base classifiers to rent. Considering that the behavior of 1 man or woman might also exchange over time, on occasion also be peculiar, we needed to use a protracted period of transit statistics to extract functions and teach classifiers.

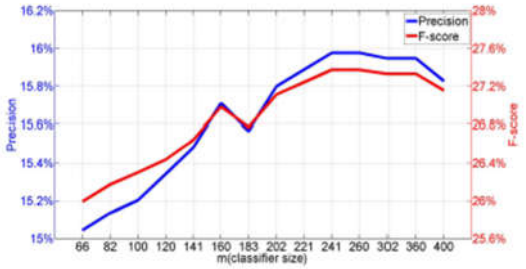


Fig. 8: Determining the number of base classifiers

In Fig. 8, we show the trend in performance as we growth the wide variety of base classifiers. The graph illustrates that each the precision and the F-score changes as the variety of base classifiers M grows from a smaller wide variety to 241. When M grows beyond 241, the performance metrics stay level or start to decrease.

D. Feature Evaluation

To in addition observe the discriminative electricity of the features, we evaluate the performance of our framework with exceptional characteristic combinations. As shown in Fig. 12, we use D, S, and H to symbolize the daily behavior, social comparison, and historical behavior features, respectively. Most extensively, the precision of the daily behavior features is improved through the social evaluation, and similarly with the aid of the historical behaviors. Such enhancements also can be discovered for metrics shown inside the different subfigures.

In Fig. 9, we additionally as compared the modeling performances on weekdays and weekends. As anticipated, on the grounds that human motilities during weekends are greater complex, the detection accuracy of our method changed into slightly decrease on weekends. For the real-time machine, we evaluated the contribution of diverse combinations of features in a similar way. The end result is shown in Fig. 16, which demonstrates that combining features helped enhance the version overall performance. In

precise, combining all three styles of features caused the pleasant overall performance in our experiments.

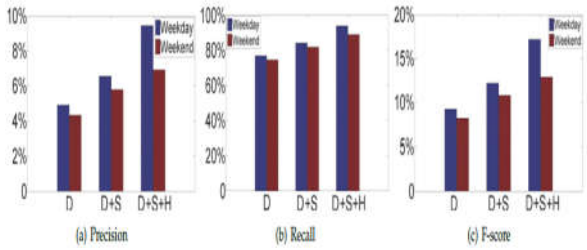


Fig. 9: The contribution of feature combinations for weekdays and weekends.

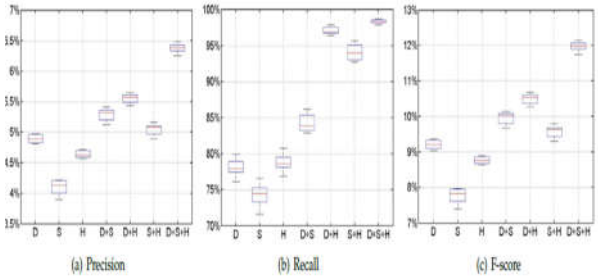


Fig. 10: The contribution of feature combinations for the real-time system.

E. Example Trajectories

We offer some example trajectories (in blue traces) projected on the map for qualitative evaluation. To make sure individuals' privacy, we selected densely populated areas so that no person changed into uniquely identifiable. We also placed the real place points to nearby points each time we could to blur the real locations visited. Every numbered box in a parent represents a station wherein the man or woman left a vehicle or boarded every other one. The numbers within the bins are series IDs.

An example of a confirmed thief's one-day trajectory is shown in Fig. 11. With all activities proven, we ought to honestly see that *Pingguoyuan* and *Shijingshan* subway stations

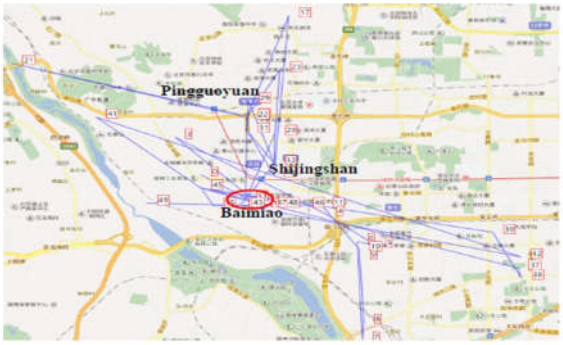
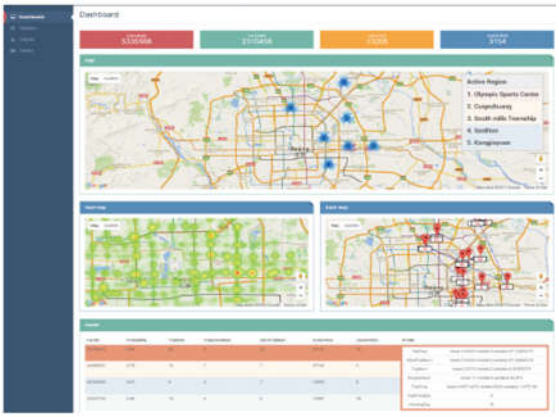


Fig. 11: The trajectory of a thief.

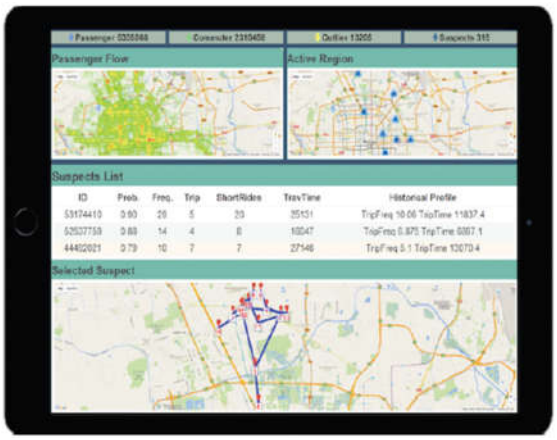
were collecting areas for this specific thief. This was an critical discovery, as it advised detectives which regions had been hubs for thieves and what number of suspects were active in that region. To recognize who's the next goal, it is important to first provide a song of a thief in the energetic regions so as to higher permit detectives' capability to catch thieves at the crime scene. Fig. 11 additionally reveals strong mobility patterns in thief activity.

VI. A PROTOTYPE SYSTEM

With the automated characteristic extraction and two-step suspect detection model, we evolved a decision aid machine for protection employees to easily spot pickpocket hotspots and recognize suspects on the crime scenes effectively. In particular, this prototype device became applied the use of bootstrap5, Java, and SparkQL. Fig. 12 is a screenshot of the graphical consumer interface (GUI), which can be considered on a computing terminal or a cell tool. The GUI has the following five simple additives, which allow customers to view suspect analytics at special degrees of element.



(a) Computing Terminal



(b) Mobile Device

Fig. 12: Screenshots of the prototype system.

Information. Precis records about the transit gadget fame are supplied on the pinnacle of the display screen, which include the total numbers of passengers, commuters, outliers, and suspects, respectively. Via diverse settings, the person is authorized to specify the time window for these data in terms of the wide variety of days.

Passenger Flows. The density of passenger flows has a high correlation with pickpocket activities. The live nation of passenger flow is proven with a warmth map, in which the density of passenger flow of each station is expressed by way of mixing the color among inexperienced and crimson. (Redder lines indicate higher densities.) This map visually identifies better trafficked regions that could be extra vulnerable to theft.

lively regions. Energetic areas of suspects at the city degree is visualized inside the "lively regions" map. Indicated through blue flashing circles, those regions are determined by means of calculating the centroids of a DBSCAN algorithm. by way of zooming in, the consumer can look at a selected location.

A. Examples of Passenger Behaviors

As stated inside the "passenger drift" function above, we visualize the passenger movement patterns on the city map to investigate the behaviors of various forms of passengers. Presents examples of consultant movement styles in exceptional passenger corporations in Beijing on a standard day from 8:00 a.m. to 11:00 a.m. every curve in the parent represents the transition between a couple of beginning and vacation spot regions, and the colour represents the visitors density (purple=high, inexperienced=low).

This provides a chicken's-eye view of the maximum dense site visitors on the city stage. We will see that the Huilongguan region, Tiantongyuan area, navy Museum, CBD place, and the Dongdan place have the highest densities. We can study several different collecting regions, inclusive of the Wudaokou region, Olympic Park region, and Beijing West Railway Station. Due to the fact that passenger flows are blended, sure unique styles are difficult to find out, particularly for travel anomalies. After making use of our technique, such patterns, which usually gift remarkably distinctive functions, can be found out. For instance, we ought to classify passengers via essential classes of purposeful regions they visit.

VII. CONCLUSION

In this paper, we developed a suspect detection and tracking system by mining large-scale transit records. The system assists in identifying pickpocket suspects' and enables active surveillance in high-risk areas. Specifically, we first constructed a feature representation for profiling passengers. Then, we established a novel two-step framework to distinguish regular passengers from pickpocket suspects. Finally, we leveraged real-world datasets from multiple sources for model training and validation, and implemented a prototype system for end users. Experimental results on real-world data showed the effectiveness of our proposed approach.

REFERENCES

[1] Paul Bouman, Evelien Van der Hurk, Leo Kroon, Ting Li, and Peter Vervest. Detecting activity patterns from smart card data. In BNAIC, 2013.

[2] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and J¨org Sander. LOF: Identifying density-based local outliers. SIGMOD Rec., 29(2):93–104, May 2000.

[3] Irina Ceapa, Chris Smith, and Licia Capra. Avoiding the crowds: understanding tube station congestion patterns from trip data. In UrbComp, pages 134–141, 2012.

[4] Olivier Chapelle, Eren Manavoglu, and Romer Rosales. Simple and scalable response prediction for display advertising. ACM Trans. Intell. Syst. Technol., 5(4):61:1– 61:34, 2014.

[5] Chao Chen, Daqing Zhang, Zhi-Hua Zhou, Nan Li, Tulin Atmaca, and Shijian Li. B-planner: Night bus ¨ route planning using large-scale taxi gps traces. In PerCom, pages 225–233, 2013.

[6] Ticiana L. Coelho da Silva, Jos´e A. F. de Macˆedo, and Marco A. Casanova. Discovering frequent mobility patterns on moving object data. In MobiGIS, pages 60–67, 2014.

[7] Marcus Felson and Ronald V Clarke. Opportunity makes the thief: Practical theory for crime prevention. Report 98, Policing and Reducing Crime Unit: Police Research Series, 1998.

[8] Yong Ge, Hui Xiong, Chuanren Liu, and Zhi-Hua Zhou. A taxi driving fraud detection system. In ICDM, pages 181–190, 2011.

[9] Marta C Gonzalez, Cesar A Hidalgo, and Albert-L´aszl´o Barab´asi. Understanding individual human mobility patterns. Nature, 453(7196):779–782, 2008.

[10] Liang Hong, Yu Zheng, Duncan Yung, Jingbo Shang, and Lei Zou. Detecting urban black holes based on human mobility data. In GIS, 2015.

[11] Shan Jiang, Joseph Ferreira Jr, and Marta C Gonzalez. Discovering urban spatial-temporal structure from human activity patterns. In UrbComp, pages 95–102, 2012.

[12] Chuanren Liu, Fei Wang, Jianying Hu, and Hui Xiong. Temporal phenotyping from longitudinal electronic health records: A graph based framework. In KDD, pages 705–714. ACM, 2015.

[13] Chuanren Liu, Hui Xiong, Yong Ge, Wei Geng, and Matt Perkins. A stochastic model for context-aware anomaly detection in indoor location traces. In ICDM, pages 449–458, 2012.

[14] Chuanren Liu, Kai Zhang, Hui Xiong, Guofei Jiang, and Qiang Yang. Temporal skeletonization on sequential data: Patterns, categorization, and visualization. IEEE Transactions on Knowledge and Data Engineering, 28(1):211–223, Jan 2016.

[15] Wei Liu, Yu Zheng, Sanjay Chawla, Jing Yuan, and Xie Xing. Discovering spatio-temporal causal interactions in traffic data streams. In KDD, pages 1010–1018, 2011.

[16] Xianglong Liu, Cheng Deng, Bo Lang, Dacheng Tao, and Xuelong Li. Query-adaptive reciprocal hash tables for nearest neighbor search. IEEE Transactions on Image Processing, 25(2):907–919, 2016.

[17] Xianglong Liu, Yadong Mu, Bo Lang, and Shih-Fu Chang. Mixed image-keyword query adaptive hashing over multilabel images. ACM Transactions on Multimedia Computing, Communications, and Applications, 10(2):22:1–22:21, 2014.

[18] Yanchi Liu, Chuanren Liu, Jing Yuan, Lian Duan, Yanjie Fu, Hui Xiong, Songhua Xu, and Junjie Wu. Intelligent bus routing with heterogeneous human mobility patterns. Knowledge and Information Systems, Forthcoming (Accepted as of Feb 2016).

[19] Wuman Luo, Haoyu Tan, Lei Chen, and Lionel M Ni. Finding time period-based most frequent path in big trajectory data. In SIGMOD, pages 713–724, 2013.

[20] Xiaolei Ma, Yao-Jan Wu, Yinhai Wang, Feng Chen, and Jianfeng Liu. Mining smart card data for transit riders travel patterns. Transportation Research Part C, 36:1–12, 2013.

[21] H. Brendan McMahan, Gary Holt, D. Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, Sharat Chikkerur, Dan Liu, Martin Wattenberg, Arnar Mar Hrafnkelsson, Tom Boulos, and Jeremy Kubica. Ad click prediction: A view from the trenches. In KDD, pages 1222–1230, 2013.

[22] Catherine Morency, Martin Tr´epanier, and Bruno Agard. Analysing the variability of transit users behaviour with smart card data. In ITSC, pages 44–49, 2006.

[23] Graeme R Newman and Megan M McNally. Identity theft literature review. Report 210459, United States Department of Justice, July 2005.

[24] Bei Pan, Yu Zheng, David Wilkie, and Cyrus Shahabi. Crowd sensing of traffic anomalies based on human mobility and social media. In GIS, pages 344–353, 2013.

Authors Profile

Nidamanuri Srinu working as Assistant Professor of CSE Department in QIS College of Engineering and Technology (Autonomous), Ongole, Andhra Pradesh, India.

Voora Yogitha Lakshmi pursuing B Tech in computer science engineering from Qis college of Engineering and Technology (Autonomous & NAAC 'A' Grade), Ponduru Road, Vengamukkalapalem, Ongole, Prakasam Dist, Affiliated to Jawaharlal Nehru Technological University, Kakinada in 2016-20 respectively.

Neelisetty Rajyalakshmi pursuing B Tech in computer science engineering from Qis college of Engineering and Technology (Autonomous & NAAC 'A' Grade), Ponduru Road, Vengamukkalapalem, Ongole, Prakasam Dist, Affiliated to Jawaharlal Nehru Technological University, Kakinada in 2016-20 respectively.

Battula Manoj Kumar pursuing B Tech in computer science engineering from Qis college of Engineering and Technology (Autonomous & NAAC 'A' Grade), Ponduru Road, Vengamukkalapalem, Ongole, Prakasam Dist, Affiliated to Jawaharlal Nehru Technological University, Kakinada in 2016-20 respectively.

Jampu Sandhya pursuing B Tech in computer science engineering from Qis college of Engineering and Technology (Autonomous & NAAC 'A' Grade), Ponduru Road, Vengamukkalapalem, Ongole, Prakasam Dist, Affiliated to Jawaharlal Nehru Technological University, Kakinada in 2016-20 respectively.

Indla Sarath Naga Sai Venkatesh pursuing B Tech in computer science engineering from Qis college of Engineering and Technology (Autonomous & NAAC 'A' Grade), Ponduru Road, Vengamukkalapalem, Ongole, Prakasam Dist, Affiliated to Jawaharlal Nehru Technological University, Kakinada in 2016-20 respectively.