Estimating Crowd Flow and Crowd Density from Cellular Data for Mass Rapid Transit

Guanyao Li, Chun-Jie Chen, Wen-Chih Peng and Chih-Wei Yi
Department of Computer Science
National Chiao Tung University, Hsinchu, Taiwan
{gli,cjchen10167,wcpeng,yi}@cs.nctu.edu.tw

ABSTRACT
Mass rapid transit (MRT) is playing an increasingly important role in many cities due to its large carrying capacity, high speed and punctuality. Understanding the crowd flow and crowd density for MRT is crucial for smart city and urban planning. The traditional way to achieve this goal is by using smart card data. However, we can only know the number of passengers entering or exiting the station from smart card data. Travel routes and transfer activity between lines of users are not captured by smart card data. Nowadays, each user has his/her own mobile phones and from the cellular data of mobile phone service providers, it is possible to know the users’ transportation mode and the fine-grained crowd flows. As such, given a set of cellular data, we aim to estimate the crowd flow of MRT passengers and crowd density of stations as well as routes. In this work, we firstly propose an efficient and scalable approach to detect MRT trips with a pre-defined reference system. Then based on the detection result, we estimate the crowd flow and crowd density by grouping and counting the MRT trips. Extensive experiments are conducted to evaluate the detection and estimation approaches on a real dataset from Chunghwa Telecom, which is the largest telecommunication company in Taiwan. The results confirmed that our approaches are suitable for MRT trips detection, crowd flow and crowd density estimation. Finally, we provide case studies to present some applications and demonstrate the usefulness of our approaches.

KEYWORDS
crowd flow and crowd density estimation, urban computing, cellular data

1 INTRODUCTION
Mass rapid transit (MRT) is playing an increasingly important role in many cities due to its large carrying capacity, high speed and punctuality. Understanding the crowd flow and crowd density for MRT is essential for urban planning, public transport network planning, public transport timetable arrangement, and route recommendation. What’s more, understanding the crowd flow and crowd density helps the telecommunication company to determine where to deploy additional cell towers in order to provide better service. In this paper, we focus on estimating crowd flow and crowd density for MRT.

In the previous studies, users’ smart card data, which record when and where users enter or exit an MRT station, have been used to estimate the MRT passenger flow[5]. However, the smart card only records the origin and destination stations of a user. It does not capture users’ travel routes and transfer activity between lines[4]. Whether a user changes lines during the trip, and when and where the change occurs are usually unknown. We cannot estimate the exact passenger number for those transfer stations, since the number of transfer passengers is unknown. For example, there are several MRT lines in Figure 1 and A, B, C, D and E are MRT stations. If a passenger takes the MRT from station A to station D, she can choose the route (A, B, C, D) in which she changes from the blue line to the red line at station C or she can choose the route (A, B, E, D) in which she changes from the blue line to the green line at station B. In both cases, the smart card system only records the original station A and destination station D but it cannot identify which transfer station the passenger chooses.

To deal with the limitation of using smart card data, we use cellular data instead. Nowadays, each user has his/her own mobile phones. When users use their mobile phones to call, send messages or access the internet, the phones are connected to a nearby cell tower. Even if the user does not use the phone, it will still connect to a nearby cell tower every hour. Using the location of cell tower to denote a user’s approximate location, it is possible to know the users’ transportation mode and the fine-grained crowd flows. However, there are several issues to estimate crowd flow and crowd density from cellular data. First, the coverage of the cell tower is large and there is intersection of coverage of some cell towers. We cannot infer the exact location, the moving direction and the speed of a user from the cellular data directly. Second, the cellular data sampling rate is irregular. It depends on the strength of the signal and whether the user uses her mobile phone during the trip. There may be no data occurring when the use passes some MRT stations. Third, some MRT routes are similar to routes of other transportation modes, which is also a challenge to distinguish MRT trips and other trips.

To deal with the above issues, we firstly propose an efficient and scalable approach for detecting both indoor and outdoor MRT trips. Some cell towers are selected as reference towers to build a reference system. Based on the reference system, a matching-based approach is proposed to detect individual MRT trips. Our approach takes both spatial and temporal features into consideration. External data, including MRT route network data and MRT travel time data are introduced to our approach. Then, based on the results of our detection approach, we estimate the crowd density of stations and routes as well as the crowd flow of origin-destination stations in a route.

In summary, the major contributions of our paper are outlined as follows:
We study on a novel and fundamental problem in urban computing, i.e., estimating crowd flow and crowd density from cellular data for Mass Rapid Transit.

We propose an efficient and scalable approach to detect both indoor and outdoor MRT trips with cellular data considering the cell tower property, spatial and temporal factors.

We conduct comprehensive experiments over the data from the largest telecommunication company in Taiwan, the results demonstrate the efficiency and scalability of our approaches.

We provide case studies to show the result of crowd flow and crowd density estimation for MRT.

The remainder of this paper is organized as follows: Section 2 discusses related works. Section 3 shows the dataset and data pre-process. Section 4 presents our MRT trip detection approach. The estimation approaches are discussed in Section 5. The evaluation results of our approach are presented in Section 6. Section 7 shows case studies. Section 8 concludes this paper.

2 RELATED WORKS

Our work is about MRT trip detection as well as crowd density and crowd flow estimation. In this section, we discuss some related works.

There have been some works about transportation mode detection. The prior works [10] [16] [17] proposed methods to detect user transportation mode from user GPS data. Compared to the GPS data, there are two challenges of using mobile phone data: it is inaccurate in determining the position of a user, and the data sampling rate is irregular[2]. To deal with these issues, [2] presented methods for mapping trajectories of cell tower latitude-longitudes to transport networks. It defined stay region, extracted trajectories between stay regions as trips and then mapped the trips to the transport network. Extracting users’ stationary stay locations from cellular data as the origin and destination of trips was prevalent in the prior research, it has also been discussed in [1] and [6]. Compared to [1] [2] [6], our proposed approach does not have to detect the stay region for each user. Instead, we mapped the cell tower to the MRT station directly utilizing a pre-defined reference system. The paper [5] proposed a method to extract trips from user call detail record(CDR) data and utilized the data from the public transport smart card system to distinguish transportation mode. A probabilistic method consisting of a Hidden Markov Model and two sub-models was proposed in [15] to identify transportation modes(driving, biking and walking). Labeled data were necessary for training in [15]. In [12], users from the same origin to the same destination were clustered into three subgroups(driving, walking and public transit) according to their travel time to infer transport mode share. Other works [7] [9] utilized the signal strength to detect the transportation mode. However, the signal strength data are not available in our work. An algorithm for MRT trip detection was proposed in [4] which utilized the property that the indoor MRT stations in Singapore are served exclusively by indoor cell towers, and cell phones outside the MRT network cannot access those towers. However, the limitations of this work are obvious. The algorithm is limited to detecting indoor MRT trips. And for other cities without the exclusive property, the algorithm in [4] is not suitable.

For crowd flow and crowd density estimation, [3] built a hard-and software system to estimate the number of passengers in a vehicle. [8] presented approaches to estimate crowd density and pedestrian flows using Wi-Fi and bluetooth data. In [13], a bluetooth scan based method was proposed to detect the crowd density. In our work, we focus on using cellular data which can be obtained more easily for telecommunication company as no extra sensors or devices are needed.

3 DATASET AND DATA PRE-PROCESS

We use users’ mobile phone cellular data from Chunghwa Telecom, which is the largest telecommunication company with a market share of 38% in Taiwan. The cellular data we used are the records of cell towers the mobile phone is connected to. A mobile phone is connected to a cell tower in two cases: active network events or passive network events[4]. Active network events include calling, sending messages or accessing the internet. Passive network events include switches between network zone or after one hour of inactivity.

An example of the user mobile phone data is shown in Table 1. The data consist of a user ID, the longitude and latitude of the connected cell tower, as well as the time stamp. The user ID was anonymized by the hashing process. Thus, the personal information of the user is unknown to the authors during the study.

Because there is intersection of the coverage of some cell towers, one major problem of using cellular data for mobility modeling is the oscillation problem[14]. Oscillation occurs when the mobile phone switches between cell towers very quickly instead of being connected to one cell tower. We denote two cell towers as $t_i$ and $t_j$. Assume that the cell towers connected by a mobile phone are $(t_i, t_j, t_j, t_i)$. If the connection of the mobile phone switches between $(t_i, t_j)$ and $(t_j, t_i)$ very quickly, the connection to $t_j$ can be regarded as resulting from the oscillation problem. To reduce the effect of the oscillation problem, we remove the records which resulted from the oscillation problem. In this example, the connection...
Estimating Crowd Flow and Crowd Density from Cellular Data for Mass Rapid Transit

Table 1: Overview of the Dataset

<table>
<thead>
<tr>
<th>User ID</th>
<th>Time stamp</th>
<th>longitude</th>
<th>latitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>-87556096</td>
<td>00:59:19</td>
<td>121.587</td>
<td>25.048</td>
</tr>
<tr>
<td>-87556096</td>
<td>00:59:20</td>
<td>121.59</td>
<td>25.04</td>
</tr>
<tr>
<td>-87556096</td>
<td>00:59:21</td>
<td>121.587</td>
<td>25.048</td>
</tr>
<tr>
<td>-87556096</td>
<td>01:59:23</td>
<td>121.587</td>
<td>25.048</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>-87556096</td>
<td>16:02:01</td>
<td>121.5</td>
<td>25.041</td>
</tr>
<tr>
<td>-87556096</td>
<td>16:02:06</td>
<td>121.5</td>
<td>25.041</td>
</tr>
</tbody>
</table>

Table 2: Example of processed data

<table>
<thead>
<tr>
<th>User ID</th>
<th>Start time</th>
<th>End time</th>
<th>longitude</th>
<th>latitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>-87556096</td>
<td>00:59:19</td>
<td>01:59:23</td>
<td>121.587</td>
<td>25.048</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>-87556096</td>
<td>16:02:01</td>
<td>16:02:06</td>
<td>121.5</td>
<td>25.041</td>
</tr>
</tbody>
</table>

to \( t_j \) will be removed in our data pre-process. After that, if two consecutive records are the same cell tower, the two records will be merged. An example of the pre-process result of Table 1 is presented in Table 2. The second record in Table 1 is removed as oscillation data and three records of the cell tower \((121.587, 25.048)\) and two records of the cell tower \((121.5, 25.041)\) in Table 1 are merged as one record respectively in Table 2.

After data pre-process, we define each record as \( \ell = (u, t_c, t_s, t_e) \), where \( u \) is the user ID, \( t_c \) is a cell tower, \( t_s \) is the start time and \( t_e \) is the end time. Then the user data can be denoted as a sequence of data records: \( L = (\ell_1, \ell_2, ..., \ell_i, ..., \ell_n) \).

4 MRT TRIP DETECTION

The MRT trip detection algorithm plays the key role in our work. The intuitive idea for detecting MRT trips from a user’s cellular data is to compare the similarity of the user’s trajectory and the MRT route. However, it is not efficient since we have to calculate the distance between each MRT station and each cell tower in the user data.

An observation is showed in the Figure 3. From the observation, we know that the mobile phone of a MRT passenger is usually connected to a cell tower near the MRT station. Utilizing this property, we propose an efficient approach in this work. The overview of the approach is presented in Figure 2. Given the cellular data of the user, the detection approach will output the detail MRT routes of the user. The detection approach consists of three sub-approaches. To deal with the issue of large coverage and intersection of coverage of cell towers, we propose a tower-station matching approach with a reference system, considering the cell tower property and spatial factors. The time interval grouping approach that utilizes the external data and temporal factors is proposed to solve the problem that different transportation modes share similar routes. To deal with the irregularity of cellular data sampling rate, we propose a station complement approach to obtain detail routes of users.

4.1 Reference System Building

To detect MRT passengers, our idea is to detect whether the user’s mobile phone connects to a cell tower near MRT stations. We use the reference system consisting of reference towers of each station to achieve this goal efficiently. The idea of the reference system is inspired by the work[11]. The reference towers are cell towers near MRT stations and to which an MRT passenger’s mobile phone will be connected with high possibility. In the example showed in the Figure 3, cell towers can be used as reference towers for their closest MRT stations respectively.

To build the reference system, we select some cell towers as reference towers for each MRT station. Most MRT stations are large and there are several gates; we therefore selected reference...
towers for each gate when building the reference system. We define two types of reference systems in our work: the $K$ Nearest Tower reference system (KNT) and $D$ meters Coverage Tower reference system (DCT).

4.1.1 KNT Reference System. A mobile phone is usually connected to the nearest cell tower with the highest probability. But if two cell towers are close to each other, there is intersection of their coverage. In this case, even in the same position, different mobile phones may be connected to different cell towers. Thus, for the KNT reference system, we select the $k$ nearest cell towers as the reference towers for each gate of the station. If a cell tower is the $K$ nearest cell tower for more than one MRT station, we regard it as the reference tower for the station to which it is closest so that each reference tower in the reference system only serves one MRT station.

4.1.2 DCT Reference System. The radius of the cell tower’s coverage is limited. It ranges from several hundred in urban areas to several thousand meters in the suburbs[2]. The closer to the cell tower, the higher the possibility that the cell phone will be connected to it. In the DCT reference system, we consider the radius of the coverage. If a station is in the $D$ meters coverage of a cell tower, the cell tower will be selected as the reference tower for the station. To make sure that each reference tower only serves one station, if there is more than one station within its $D$ meters coverage, we define the tower as the reference tower of the closest station.

Figure 4 shows the overview of all the cell towers in Taipei (the green point), the gates of MRT stations (the red point) and the reference towers of each station with different reference systems (the blue point). Compared to the non-reference towers, the reference towers are very close to the corresponding MRT stations. It illustrates that the reference towers can be used as the distinction to identify the MRT trips and non-MRT trips.

4.2 Tower-Station Matching

After the reference system is built, we can use these cell towers as a distinction between MRT trips and non-MRT trips. In the user’s data, if a connected cell tower $\ell_1, \ell_2$ is a reference tower of MRT station $s_i$, the record $\ell_1$ will be matched to the station $s_i$.

For example, given the reference system showed in Table 3 and the user’s data $(\ell_0, \ell_1, \ell_2, \ldots, \ell_9)$, if the cell tower connected record $(\ell_0, \ell_1, \ell_2, \ldots, \ell_9)$ is $(\ell_0, \ell_1, \ell_2, \ell_3, \ell_4, \ell_5, \ell_6, \ell_7, \ell_8, \ell_9)$, the raw data are matched to be $(s_1, s_2, s_3, s_4)$ since $\ell_1, \ell_2, \ell_3, \ell_4$ are reference towers for $s_1, s_2, s_3, s_4$ respectively.

4.3 Time Interval Grouping

In the tower-station matching step, we detect the MRT trip from the spatial dimension and identify some potential MRT stations. But only considering the spatial dimension is not enough. If the user was driving a car past the MRT stations or taking a bus with a similar route, her mobile phone may also be connected to some reference towers. However, the travel time from one station to another is distinct for different transport modes. And the speed of MRT is higher than the speed of other public transportation modes and cars. In this step, we take the temporal factors into consideration and use the travel time as a distinction to detect trips.

Given the station sequence obtained from the last step, we group the station to a trip based on the time interval and the real travel time between every two consecutive stations in the station sequence. For two consecutive stations $s_i$ and $s_{i+1}$, we denote the time interval of the two stations as $TI(s_i, s_{i+1})$ and the real travel time as $TT(s_i, s_{i+1})$. If $TT(s_i, s_{i+1}) < TI(s_i, s_{i+1}) + \theta$, $(s_i, s_{i+1})$ is regarded as an MRT trip. $\theta$ is set to be 60s in our work. Otherwise, $s_i$ is regarded as the destination station of the last MRT trip while $s_{i+1}$ is regarded as the origin station of the new MRT trip. The real travel time from one station to another is available at DataTaipei[1], which is an open data platform of the Taipei government. The time interval of two consecutive stations in the station sequence is calculated as follows: if $\ell_1, \ell_2$ and $\ell_{i+1}, \ell_{i+2}$ is the reference tower of $s_i$ and $s_{i+1}$ respectively, then, $TI(s_i, s_{i+1}) = \ell_{i+1} - \ell_i, \ell_{i+2}$.

We also consider the connection duration for grouping. If the duration of staying in a station $s$ is more than $\beta$, the station $s$ is regarded as the destination station of the last MRT trip and also the origin station of the new MRT trip. $\beta$ is set to be 30min in the experiment. The duration of staying in a station is the duration of being connected to the reference tower of the station.

We continue the example in Section 4.2. If $TI(s_1, s_2) < TI(s_1, s_2) + \theta$, $TI(s_2, s_3) < TT(s_2, s_3) + \theta$ while $TI(s_3, s_4) > TT(s_3, s_4) + \theta$, then $(s_1, s_2, s_3)$ and $(s_4)$ are detected as two candidate MRT trips.

4.4 Station Complement

To better capture the origin station and destination station of the MRT trip, we utilize a complement method inspired by [4]. Given the candidate MRT trip $(s_1, s_2, \ldots, s_n)$, we check the connected cell tower preceding the first station $t_0$ and the connected cell tower behind the last station $t_n$ in the raw data. If the distance from $t_0$ to an MRT station $s$ is smaller than a distance threshold $\Delta d$ and the time interval is smaller than $TT(s, s_1) + \Delta t'$, we extend the MRT trip and regard $s$ as the origin station instead. For the destination station complement, we adopt the same strategy. $\Delta d$ is set to be 50m and $\Delta t'$ is set to be 10min.

In the above example, given the candidate trip $(s_1, s_2, s_3)$ and the raw data $(t_0, t_1, t_2, t_3, t_4, t_5, t_6)$, the connected cell tower preceding the reference tower of $s_1$ is $t_0$. If the distance from $t_0$ to the station $s_1$ is smaller than $\Delta d$ and the difference between the time interval and travel time satisfies the time threshold, the origin station of the MRT trip is updated to be $s_1$. The connected tower after the reference tower of $s_3$ is $t_7$. If the distance from $t_7$ to the station $s_3$ is smaller than $\Delta d$ and the time interval satisfies the time constraint, the destination station of the MRT trip is updated to be $s_3$. Assume that the trip $(s_4)$ does not change after complement. In our work, we define that each trip consists of at least two stations. Thus, the exact MRT trip is $(s_5, s_1, s_2, s_3, s_6)$.

---

After that, we complement the MRT trip according to the MRT network. If two consecutive stations $s_i$ and $s_j$ obtained in the last step are not consecutive stations in the MRT network, we insert the stations between $s_i$ and $s_j$ in the MRT network into the trip. If there are multiple routes from $s_i$ to $s_j$ in the MRT network, we choose the route whose travel time is closest to the time interval. For those stations complemented in this step, we estimate the time stamp of the station according to the real travel time from the station with a time stamp.

5 CROWD FLOW AND CROWD DENSITY ESTIMATION

Based on the results of our proposed MRT trip detection approach, we define three kinds of passengers:

- (In-Passenger) The passenger who enters and departs from the same station.
- (Out-Passenger) The passenger who arrives at the station by MRT and goes out of the station.
- (Transfer Passenger) The passenger who arrives at the station by MRT, changes the MRT line and departs from the station as the transfer station.

The smart card system only records the origin and destination stations, so we can only estimate the number of in-passengers and out-passengers from the smart card system data. However, since we can obtain the exact MRT routes of a user with our proposed detection approach, we can estimate the number of the three kinds of passengers based on the detection result. In this paper, we focus on estimating the crowd density of a station, the crowd density of different lines at a transfer station and the crowd flow of origin-destination stations in a route.

The crowd density of a station is the sum of the number of the three kinds of passengers. We obtain the number of the trips whose origin, destination or transfer station is the target station and the line is the target line.

The crowd flow of origin-destination stations in a route is the number of passengers from the given origin station to the given destination station by the route. Understanding the crowd flow between any origin-destination stations is essential. It is helpful for route recommendation, route arrangement and schedule arrangement. To estimate the crowd flow, we count the trips that from the origin station to the destination station by the route among all trips in the detection result.

Some case studies about the crowd density and crowd flow estimation will be presented and discussed in the Section 7.

6 EVALUATION

To evaluate the performance of our approach, we conducted our experiments in three parts. In the first part, we evaluated the accuracy of our detection and estimation approaches. In the second part, we discussed the precision and recall of our detection approach. Finally, we tested the efficiency and scalability of our detection approach.

6.1 Performance of the estimation approach

To evaluate the accuracy of our detection and estimation approaches, we estimated the in-passenger and out-passengers number of two stations based on the trip detection result and compared the estimation results with the ground truth released by the Taipei Metro Company. The Pearson’s correlation is used as metric for evaluation.

We applied our approach to the data of 10% of Chunghwa Telecom users in Taipei from 2017/01/06 to 2017/01/12. The number of users is around 310,000. In the KNT reference system, $K$ was set to be 1, 2 and 3; and $D$ was set to be 100m, 200m and 300m in the DTC reference system. The MRT stations we selected for evaluation are the Nanjing Fuxing Station and the Taipei 101 Station, which are two of the most important stations in Taipei. To estimate the number of in-passengers and out-passengers, we firstly detected the MRT trips of all users. Then we counted the trips according to the origin and destination at hourly intervals. The market share of Chunghwa Telecom is 38% and the data we used were 10% of
Table 4: Correlation between the estimation result and ground truth

<table>
<thead>
<tr>
<th>Type</th>
<th>Correlation</th>
<th>Type</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>K = 1</td>
<td>0.956394</td>
<td>D = 100</td>
<td>0.954600</td>
</tr>
<tr>
<td>K = 2</td>
<td>0.956323</td>
<td>D = 200</td>
<td>0.956437</td>
</tr>
<tr>
<td>K = 3</td>
<td>0.956169</td>
<td>D = 300</td>
<td>0.960197</td>
</tr>
</tbody>
</table>

Table 5: Evaluation of the detection result

<table>
<thead>
<tr>
<th>Type</th>
<th>Precision</th>
<th>Recall</th>
<th>F-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>K = 1</td>
<td>0.9144</td>
<td>0.7565</td>
<td>0.8280</td>
</tr>
<tr>
<td>K = 2</td>
<td>0.7858</td>
<td>0.7678</td>
<td>0.7767</td>
</tr>
<tr>
<td>K = 3</td>
<td>0.7680</td>
<td>0.7927</td>
<td>0.7802</td>
</tr>
<tr>
<td>D = 100</td>
<td>0.8315</td>
<td>0.7379</td>
<td>0.7819</td>
</tr>
<tr>
<td>D = 200</td>
<td>0.7629</td>
<td>0.7990</td>
<td>0.7805</td>
</tr>
<tr>
<td>D = 300</td>
<td>0.6402</td>
<td>0.8101</td>
<td>0.7152</td>
</tr>
</tbody>
</table>

6.2 Performance of the detection approach

In this part of the experiment, we used a small dataset in which the truth of the MRT trips had been labelled by the users for evaluation. The dataset consists of 7 days data of 10 users in Taipei. To validate the effectiveness of the approach, the precision, recall and F-1 score are investigated. We firstly define the precision and recall for one trip.

**Definition 6.1.** (Precision of a detected trip) Given a detected trip \( T_i \) with \( |T_i| \) stations, if the trip exists and the real trip \( T'_j \) consists of \( |T'_j| \) stations, then the precision of the trip is \( p_i = (|T_i \cap T'_j|)/|T_i| \). Otherwise, \( p_i = 0 \).

**Definition 6.2.** (Recall of a real trip) Given a real trip \( T'_j \) with \( |T'_j| \) stations, if the trip is detected and the detected trip \( T_i \) has \( |T_i| \) stations, the recall is \( r_j = (|T_i \cap T'_j|)/|T'_j| \). Otherwise, \( r_j = 0 \).

Then given the set of detected trips \( \{T_1...T_i...T_n\} \) and a set of real trips \( \{T'_1...T'_j...T'_m\} \), the precision is \( \text{Precision} = (\sum_{i=1}^{n} p_i)/n \), the recall is \( \text{Recall} = (\sum_{j=1}^{m} r_j)/m \) and the F-1 score is \( F_1 = 2 \ast (\text{precision} \ast \text{recall})/(\text{precision} + \text{recall}) \).

The precision, recall and F-1 score of our approach on the dataset is shown in Table 5. All of them confirm the good performance of our approach. In Table 5, the precision decreases when \( K \) or \( D \) becomes larger while the recall increases. That is because when more cell towers are selected as the reference towers, more trips will be detected as MRT trips. But at the same time, some of the non-MRT trips are detected as MRT trips by mistake. When considering the F-1 score, the approach with \( K = 1 \) has a better performance than other approaches.

6.3 Efficiency and Scalability

To evaluate the efficiency and scalability of our detection approach, we sampled the cellular data on 2017/01/06 with different sampling rate: 0.2, 0.4, 0.6, 0.8 and 1. We ran our proposed approach with various reference systems on the datasets of different size. The running time is shown in Figure 7. The running time of our approach increases when the size of the dataset increases. The
We utilize the KNT reference system with the DCT approach, we present the overview of the crowdedness of MRT stations and MRT routes at different times in Taipei. Then we estimate the crowd density for MRT stations and routes. We firstly proposed an efficient and scalable approach to detect MRT trips from cellular data. In our proposed detection approach, both spatial and temporal factors were considered. Then we estimated the crowd flow and crowd density in a day.

### 7.2 Crowd density of a station

We aim to show the crowd density of a station for each hour. Figure 10 shows the results for Taipei Main station, which is one of the most important and crowded MRT stations in Taipei. We use the data from 2017/01/07 to 2017/01/10. From the figures, we can learn about the estimation number of the three kinds of passengers for each hour. The estimation number of transfer-passengers helps us have a more comprehensive understanding of the crowd density of the station. In Figure 10(a), the trend of number of passengers on a workday(2017/01/09 and 2017/01/10) is different from the trend on the weekend(2017/01/07 and 2017/01/08). There are two peaks in workday’s trend, one is around 8 : 00 which is the time people go to work and one is around 18 : 00 which is the time people go home after work. But for the weekend, there is only one peak at around 17 : 00. The results in Figures 10(b) and 10(c) lead to a similar conclusion.

### 7.3 Crowd density of different lines

We take the Taipei Main Station for example in the case study. It is the intersection of the MRT blue line and MRT red line. We estimate the crowd densities of the two lines for each hour. Figure 11(a) presents the number of passengers departing from Taipei Main Station by different lines each hour on 2017/01/07(weekend) and 2017/01/10(workday). Similarly, Figure 11(b) presents the number of passengers arriving at Taipei Main Station by different lines for each hour. From the result, we learn that the Blue line is busier than the Red line at the Taipei Main Station.

### 7.4 Crowd flow of Origin-Destination Stations

Given the origin station and destination station, we estimate the crowd flow from the origin station to the destination at hourly intervals. Figure 8 shows the number of all passengers(AP) from Minquan W.Rd. Station to Taipei Main Station as well as the transfer passengers(TP) who are from Minquan W.Rd. Station and then change their MRT line at Taipei Main Station. The figure reveals that the trends are different on the weekend(2017/01/07) and on workdays(2017/01/10). Both the peak time as well as the number of transfer passengers and all passengers differ between the weekend and workdays.

### 8 CONCLUSION

In this paper, we aimed to estimate the crowd flow and crowd density for MRT stations and routes. We firstly proposed an efficient and scalable approach to detect MRT trips from cellular data. In our proposed detection approach, both spatial and temporal factors were considered. Then we estimated the crowd flow and crowd density in a day.
density of stations and routes based on the MRT trip detection result. We conducted extensive experiments to show the effectiveness, efficiency and scalability of our approaches on the data from Chunghwa Telecom, which is the largest telecommunication company in Taiwan. We also provided several case studies. The case study demonstrated that we could have a more comprehensive understanding of the crowd flow and crowd density for MRT utilizing our proposed detection and estimation approaches.
REFERENCES


