Spatio-temporal Activity Recognition Technology to Achieve Proactive Navigation

Hiroyuki Toda, Shuhei Yamamoto, and Takuya Nishimura

Abstract

We are developing spatio-temporal data analysis technologies to achieve proactive navigation that supports user activities by taking changes in the user's situation and in the surrounding environment into account. We introduce here spatio-temporal activity recognition technologies, which are key elements of spatio-temporal data analysis. The recognition technologies understand the user's activities from spatio-temporal movement data (moving trajectories, acceleration of movement, etc.) recorded by mobile devices.

Keywords: spatial trajectories, spatio-temporal data, data mining

1. Introduction

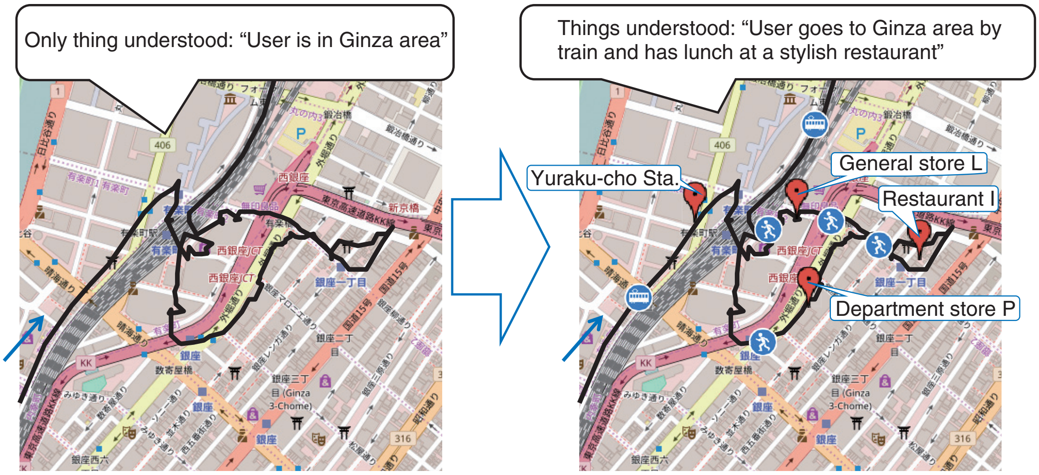
The word *navigation* suggests the idea of a routing assistance service such as a car navigation system that assists users to get from a starting location to a destination. The goal of our proactive navigation includes but is not restricted to routing assistance. For example, it provides users with suggestions for that day's dinner plans on the basis of their past activity histories, and it provides information to alert them of locations where traffic accidents or near misses frequently occur. In summary, the goal of our proactive navigation is to provide useful information for supporting users' spatio-temporal activities on the basis of past and present situations, as well as future situations in which changes may occur with respect to users and their surroundings [1].

2. Spatio-temporal activity recognition technologies

There has been a surge in popularity recently for mobile devices that can record people's activities using sensors. For example, about 1.36 billion smartphones with GPS (global positioning system) sensors were shipped around the world in 2016, a good indication that the market for these devices is continuing to grow [2]. Various other devices that use sensors are also coming into widespread use. These include wearable devices such as smart watches or smart bands that can sense vital data and motion data, and dashboard cameras (dashcams) that can record driving situations. These devices make it possible to record a wide variety of people's activities.

However, what these devices record is basically time-series numerical data. For example, moving trajectories are the sequence of the combination of timestamp, latitude, and longitude. When we draw a moving trajectory based on the latitude and longitude on a map, people who know the area can understand the outline of the moving behavior, but those who do not know the area understand only that the person went to that area (**Fig. 1(a)**).

We have developed spatio-temporal activity recognition technologies to tackle this problem. Using the technologies to analyze moving trajectories enables us to understand not only the areas that the users visited, but also the shops or places they visited [3] and


Only thing understood: “User is in Ginza area”
Things understood: “User goes to Ginza area by train and has lunch at a stylish restaurant”
Yuraku-cho Sta.
General store L
Restaurant I
Department store P

(a) Moving trajectory

(b) Analysis results in recognizing spatio-temporal activity

Map data: ©OpenStreetMap Contributors

Fig. 1. Examples of moving trajectory and spatio-temporal activity recognition results.

the transportation mode they used to get there (**Fig. 1(b)**) [4]. This knowledge makes it possible for us to select suitable information for users.

In the following section we describe two state-of-the-art techniques employed in our spatio-temporal activity recognition technologies. One is a destination prediction technique that can predict the destination of a moving user. The other is a travel condition prediction technique that can detect travel conditions such as slow-moving traffic jams or near miss incidents while driving, through the use of time-series videos and sensor data sequences such as dashcam data.

3. Destination prediction technique

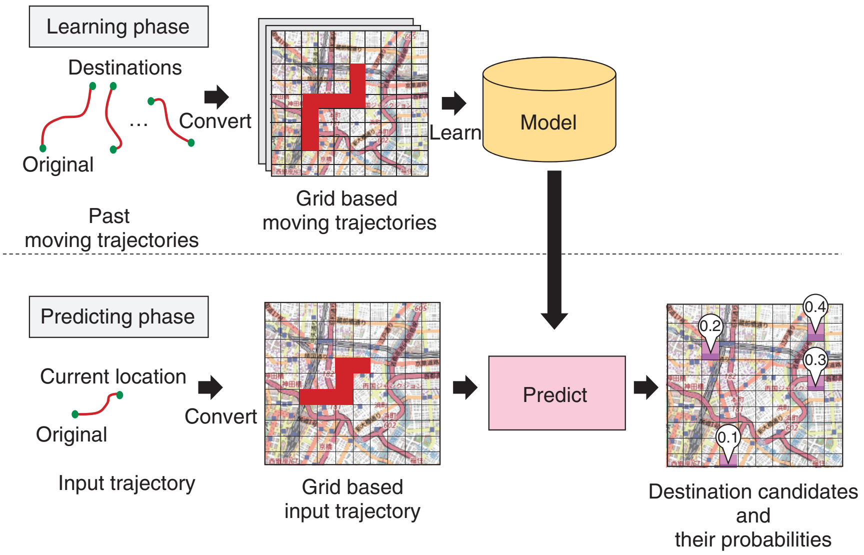
This method consists of two phases: a learning phase and a prediction phase. In the learning phase, a user model is constructed that memorizes a user’s movement trends from the user’s past moving trajectories. In the prediction phase, the destination candidates are predicted using the learned model and the user’s movement data from the original location to the current one (**Fig. 2**) [5].

Two requirements need to be simultaneously satisfied to achieve accurate prediction. One is that the long-term dependency of the movement from an original location to the current one must be taken into account (1). The other is that the data sparsity problem must be factored in (2). However, although the two requirements are related, a simple solution for requirement (1) does not satisfy requirement (2).

To satisfy both requirements, we propose the use of a recurrent neural network (RNN)[*1] for predicting destinations. An RNN is a neural network for modeling sequence data. Specifically, we regard moving trajectories as transitions on a grid space (**Fig. 3(a)**), and the transitions are modeled by the RNN. The RNN model we utilize can use long-term dependency to achieve accurate prediction when the current transitions match the past transition data, and can predict destinations on the basis of only the latest transitions even when the current transitions from the start location to the current one do not match the past transition data. Thus, this method satisfies the two requirements.

However, this RNN based model has a huge computational cost when we use it for prediction. This is because in the simplest terms, it predicts one transition with one calculation. It therefore needs G^M times calculation when the number of grids is G (several hundred or more), and the number of steps to the destination is M (several dozen or more). These are not realistic numbers for calculation purposes. To solve this problem, we use a sampling simulation

[*1] RNN: A type of neural network that mainly uses modeling of sequence data.

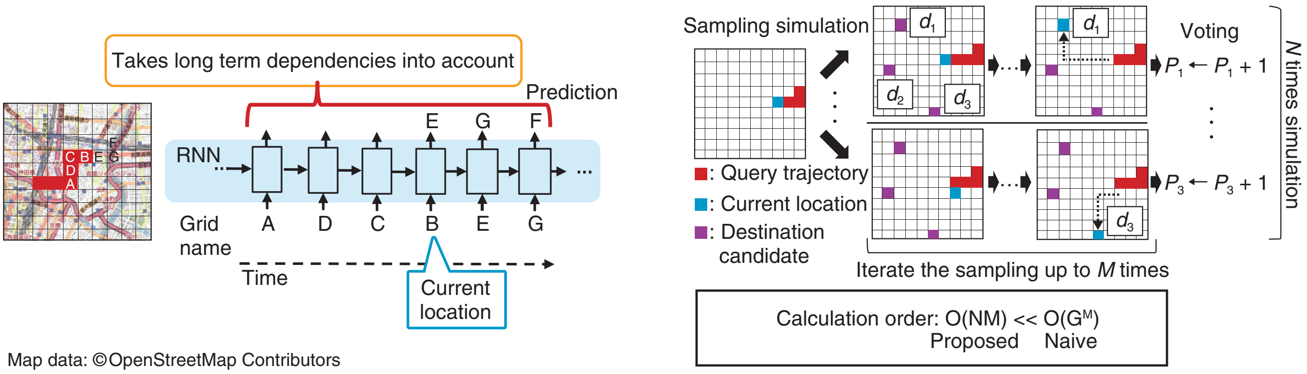


Map data: ©OpenStreetMap Contributors

Fig. 2. Outline of destination prediction technique.

(a) Method regards moving trajectories as transitions on grid space and models transition via an RNN

(b) Reducing computational cost by sampling simulation



Map data: ©OpenStreetMap Contributors

G: Number of grids (several hundred or more)
M: Number of steps to destination (several dozen or more)
N: Number of simulations (about 100)

Fig. 3. Technical point of destination prediction technique.

(**Fig. 3(b)**) to reduce the number of calculations needed. The results we obtained confirmed that our method can predict destinations in about one second by using a conventional personal computer server environment.

When this technique is used to analyze the personal trajectories of smartphone users, it predicts the users' future movements and provides the users with useful information about future destination candidates. When it is used in a car navigation system, it provides traffic information about the route to the destination, and information about alternative routes on the basis of the traffic information even if the destination is not set.

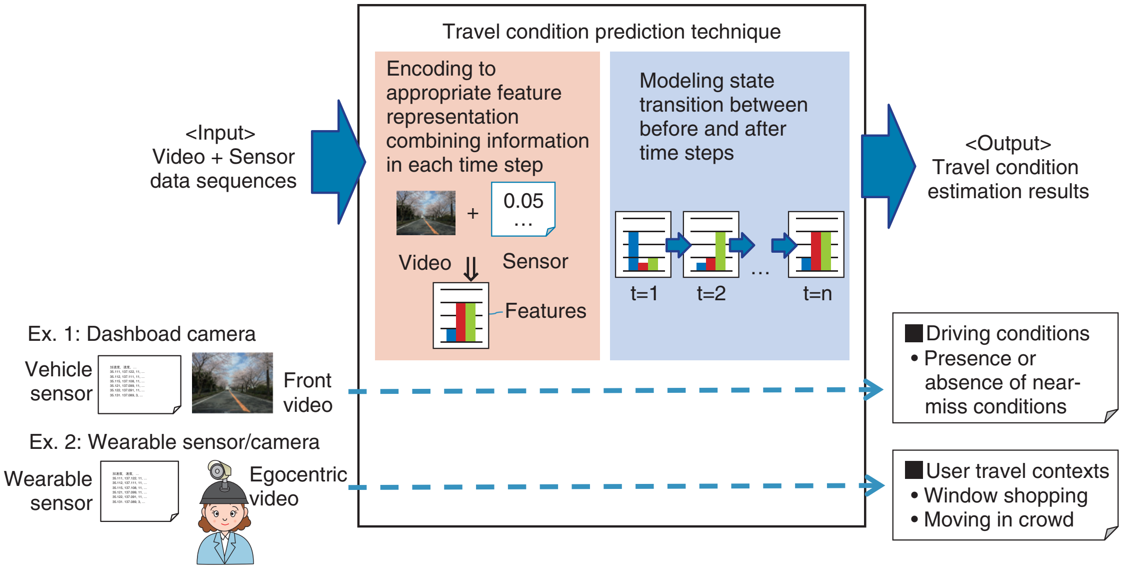

Travel condition prediction technique
Encoding to appropriate feature representation combining information in each time step
0.05
…
Video
Sensor
Features
Modeling state transition between before and after time steps
t=1
t=2
…
t=n
<Input>
Video + Sensor data sequences
<Output>
Travel condition estimation results
Ex. 1: Dashboard camera
Vehicle sensor
Front video
Ex. 2: Wearable sensor/camera
Wearable sensor
Egocentric video
Driving conditions
• Presence or absence of near-miss conditions
User travel contexts
• Window shopping
• Moving in crowd

Fig. 4. Overview of travel condition technique.

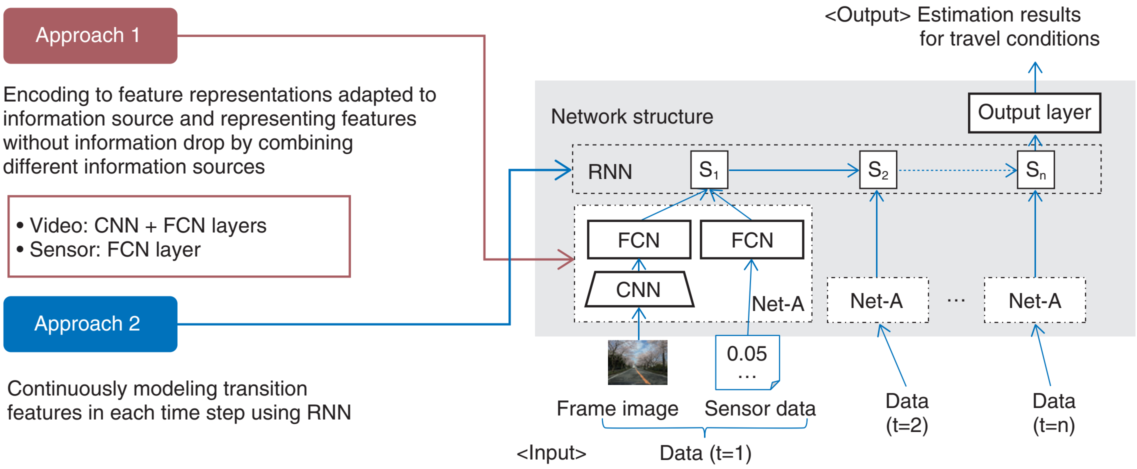

Approach 1
Encoding to feature representations adapted to information source and representing features without information drop by combining different information sources
• Video: CNN + FCN layers
• Sensor: FCN layer
Approach 2
Continuously modeling transition features in each time step using RNN
<Output> Estimation results for travel conditions
Network structure
Output layer
RNN
S1
S2
Sn
FCN
FCN
CNN
Net-A
Net-A
…
Net-A
0.05
…
Frame image
Sensor data
Data (t=2)
Data (t=n)
<Input>
Data (t=1)

Fig. 5. Neural network structure of travel condition prediction and our approaches.

4. Travel condition prediction technique

Our travel condition prediction technique estimates the moving situations of people from multimodal time-series data recorded by both video images and sensors. For instance, it automatically detects the presence or absence of near-miss scenes from a dashboard video camera in a vehicle and identifies specific user contexts such as walking in a crowd or window shopping from a wearable camera and wearable sensor (**Fig. 4**).

Two difficulties arise when using such data to accurately estimate travel conditions: extracting feature representations for travel conditions from different information sources and modeling state transitions that have time variations.

We solved these problems by using a novel neural network with our travel condition prediction technique. The novel neural network is based on an RNN that can treat state transitions of time-series data. An overview of our proposed model is shown in **Fig. 5**. First, video images and sensor data, which are input in each time step, are encoded to feature representations by different components of the neural network. A frame image in each time step obtained from video acquires the appropriate feature representation via a fully connected network (FCN) layer and a convolutional neural network (CNN)[*2] layer, which are frequently used in image analysis. Also, sensor data are encoded to suitable feature representations using the FCN layer. Feature representations are extracted by concatenating a feature representation of a frame image and that of sensor data in each time step. Modeling these features using the RNN in correct time sequences makes it possible to take state transitions with time variations into account.

In our experimental evaluation, a near-miss scene was identified by a dashboard video camera, and our travel condition prediction technique detected dangerous driving scenes with higher accuracy than that obtained with three baseline methods: without video images, without sensor data, or without state transitions with time variations, that is, using an RNN. Moreover, we confirmed that the detection accuracy was approximately 90% [6].

NTT Communications and Nippon Car Solutions are considering using our technique as a way to reduce traffic accidents [7]. We are also considering another way to use it, namely by combining location information with actual near-miss cases collected from many vehicle dashboard video cameras. This will enable us to make a *hiyari-hat* (near-miss accidents) map that can pinpoint dangerous places where near-miss situations frequently occur. Various other applications can be considered such as using wearable cameras and sensors to understand specific user contexts and to accurately identify sports scenes from videos recorded within a stadium and sensors fitted for each player.

5. Future work

We plan to collaborate with partner companies to achieve practical use of these technologies. However, we face several challenges in attempting to improve them. For example, the current destination prediction technique cannot predict places that the users have never been to, but we expect it to be able to do so in a practical manner in the future. The current travel condition prediction technique can detect the presence or absence of near misses but cannot accurately understand the types of near misses that have occurred. One future task will be to improve the accuracy and coverage range of the existing technologies.

We should also note the fact that various sensors can record wide-ranging activities through the development of the IoT (Internet of Things) environment. For this environment, we will need to develop our technologies so that they can accurately understand users' spatio-temporal behavior through the use of various sensors and in various situations. To make the proactive navigation truly effective, we must develop our technologies by utilizing understandings of past events to predictions, from known areas to unknown areas, and from simple behavior to diverse behavior.

References

[1] J. Ikedo, T. Horioka, Y. Niikura, Y. Koike, H. Sawada, and Y. Muto, "Proactive Navigation Optimized for Individual Users," NTT Technical Review, Vol. 13, No. 7, 2015.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201507fa4.html

[2] Press release issued by TrendForce on January 25, 2017, "TrendForce Reports Global Smartphone Production Volume Totaled 1.36 Billion Units; Samsung Held On as Leader While OPPO and Vivo Burst into Global Top Five."
http://press.trendforce.com/node/view/2741.html

[3] K. Nishida, H. Toda, T. Kurashima, and Y. Suhara, "Probabilistic Identification of Visited Point-of-interest for Personalized Automatic Check-in," Proc. of UbiComp2014 (the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing), pp. 631–642, Seattle, WA, USA, Sept. 2014.

[4] Y. Endo, H. Toda, K. Nishida, and A. Kawanobe, "Deep Feature Extraction from Trajectories for Transportation Mode Estimation," In Advances in Knowledge Discovery and Data Mining, Proc. of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) 2016, pp. 54–66, Auckland, New Zealand, Apr. 2016.

[5] Y. Endo, K. Nishida, H. Toda, and H. Sawada, "Predicting Destinations from Partial Trajectories Using Recurrent Neural Network," In Advances in Knowledge Discovery and Data Mining, Proc. of PAKDD 2017, pp. 160–172, Jeju, South Korea, May 2017.

[6] K. Ito, "AI Auto-detects Dangerous Driving," Image Laboratory, Vol. 28, No. 2, pp. 45–47, 2017 (in Japanese).

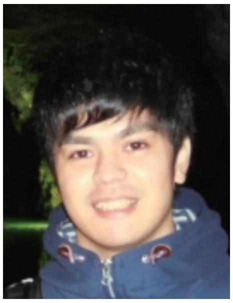
[7] Press release issued by NTT Communications on September 26, 2016 (in Japanese).
http://www.ntt.com/about-us/press-releases/news/article/2016/20160926_2.html

[*2] CNN: A type of neural network inspired by the organization of the visual cortex of humans. It is mainly used for image recognition.

Hiroyuki Toda

Senior Research Engineer, Supervisor, Proactive Navigation Project, NTT Service Evolution Laboratories.

He received a B.E. and M.E. in materials science from Nagoya University in 1997 and 1999, and a Ph.D. in computer science from University of Tsukuba in 2007. He joined NTT in 1999. His current research interests include information retrieval, data mining, and ubiquitous computing. He is a member of the Information Processing Society of Japan (IPSJ), the Institute of Electronics, Information and Communication Engineers (IEICE), the Japanese Society for Artificial Intelligence (JSAI), the Database Society of Japan (DBSJ), and the Association for Computing Machinery (ACM).



Shuhei Yamamoto

Researcher, Proactive Navigation Project, NTT Service Evolution Laboratories.

He received a B.L.S., M.I., and Ph.D. in informatics from University of Tsukuba in 2012, 2014, and 2016. He joined NTT Service Evolution Laboratories in 2016. He is currently studying data mining and analysis of real spatio-temporal data obtained from devices such as dashboard camera sensors and wearable sensors. He is a member of IPSJ and DBSJ.

Takuya Nishimura

Researcher, Proactive Navigation Project, NTT Service Evolution Laboratories.

He received a B.E. and M.E in social informatics from Kyoto University in 2012 and 2014. He joined NTT Service Evolution Laboratories in 2014. He is currently studying data mining from geographic information systems, social media, and human trajectories. He is a member of DBSJ.