Predicting Destinations with Smartphone Log using Trajectory-based HMMs

Sun-You Kim, Sung-Bae Cho
Dept. of Computer Science
Yonsei University
Seoul, Korea
sykim@sclab.yonsei.ac.kr, sbcho@yonsei.ac.kr

***Abstract*—With the spread of smartphones, it is easy to obtain sensor data from users, and location-based service (LBS) becomes the most common service in mobile industry. Predicting the user's destination can lead to a variety of services in mobile devices. In addition to the user's final destination, the locations during movement are also important in LBS. In this paper, we propose a destination prediction method based on hidden Markov models for representing the paths using sensor data from smartphone and identifies the destination and intermediate locations in future moving with visiting probabilities. In order to demonstrate the usefulness of the proposed method, we compare it with Dynamic Time Warping (DTW), a method of template matching. Experiments with the data collected by 10 college students for five months confirm that the proposed method results in 12.67 times faster and 2.88 times more accurate than the DTW.**

Keywords-destination prediction; forecasting; hidden Markov model; location-based service.

I. INTRODUCTION

The proliferation of smartphones facilitates to obtain various sensor data from the users, and a wide range of services using a variety of sensor data are introduced. Location-based service (LBS) is the most common service for utilizing the sensor information. It can extract key locations and identify the exact coordinates using user location provided by smartphone. Also, it predicts where the user moves next and provides the information required in the future in advance [1]. As a result, the technology to predict the destinations and movements of the user is required in mobile environment.

Users make a trajectory by moving locations along the flow of time. A trajectory based on the information of locations visited can be obtained. Prediction of destination is to find out the next location in future based on the information of movements until now. In other words, when the information of location movement is $Trajectory_t = \{L_1, L_2, \cdots, L_t\}$ until the current time t, the prediction of location is to find out the location L_{t+n} at time $t+n$.

In order to search L_{t+n}, we should look for a movement which has the same pattern with $Trajectory_t$ in the past [2]. As a path is a subsequence of trajectory, the path becomes a moving pattern. Path is a set of locations, each of which is constructed by temporal and spatial information.

The problem of predicting destination can be defined using the path. The past path $P_{optimal}$ with the highest similarity using the current moving path $P_{present}$ is determined, and the endpoint of path $P_{optimal}$ is said to be the destination $L_{destination}$.

In this paper, we propose a destination prediction method which utilizes hidden Markov models for representing the paths using sensor data from smartphone, estimates the destination L_{t+n} based on $P_{present}$, and finds out the intermediate locations $L \in \{L_{t+1}, L_{t+2}, \cdots, L_{t+n-1}\}$. The problem of destination prediction is shown in Fig. 1.

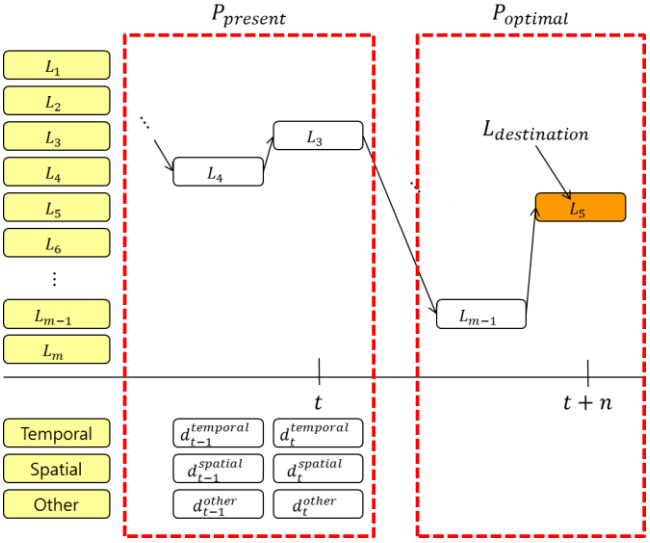

Figure 1. The problem of destination prediction

The rest of this paper is organized as follows. Section II describes the related works in destination prediction. Section III describes the proposed method. The proposed method consists of constructing model and predicting destination. Section IV addresses the result of experiments. Section V summarizes the paper and draws a conclusion.

II. RELATED WORKS

The study of destination prediction is related with comparing the information of previous visits with the current location of visit along the flow of time, in order to identify the destination. The trajectory of locations along the flow of time can be expressed as time series data. There are three types of classification methods for time series data. First, feature-based classification finds the decision boundary. Second, the sequence distance-based classification method classifies using the class of closest distance of time-series data. Third, the

 ISBN: 978-1-61208-366-7

model-based classification method identifies the most probable class after creating the probability model [3].

TABLE I. PREVIOUS STUDIES ON DESTINATION PREDICTION

Year	Author	Data	Method	Method category
2013	Do, et al.	Apps, Call history, Bluetooth	Linear regression	Feature-based
2012	Lu, et al.	GPS, Bluetooth	SVM	Feature-based
2012	Mathew, et al.	GPS	HMM	Model-based
2012	Gambs, et al.	GPS	Mobility Markov chain	Model-based
2011	Kim, et al.	GPS	Bayesian network	Feature-based
2009	Monreale, et al.	GPS	Trajectory pattern tree	Feature-based
2009	Lee, et al.	GPS	DTW	Sequence distance-based
2008	Burbey, et al.	GPS	PPM	Model-based
2007	Akoush, et al.	Cell ID, Cell history, time	Bayesian neural network	Feature-based

A. *Feature-based classification*

The studies which incorporated feature-based classification classified the next destination by using the context information of the current state and the current location. Do, et al. proposed a destination prediction method based on linear regression using location, Apps, call history, and Bluetooth [4]. Lu predicted destinations using SVM with place, Bluetooth, WLAN, and call history as inputs [5]. Monreale, et al. predicted a destination using the method of Trajectory Pattern Tree that uses GPS trajectory [6]. Also, Akoush, et al. performed a destination prediction by applying Bayesian neural networks with cell ID, cell history, and time as variables [7]. In addition, Kim, et al. predicted the user's destination using a Bayesian network created from history information of visited locations in the past [8]. However, the studies using feature-based classification method predicted the next location with fragmented information only. Moreover, because it does not consider the path, it is impossible to know the location information of the intermediate paths.

B. *Sequence distance-based classification*

Sequence distance-based classification, with the use of time series data, is a classification method which works by determining the class which has the highest similarity among the stored templates. Lee, et al. classified the user's destination by using dynamic time warping, which is a method of determining the similarity of the pattern of the two GPS paths [9]. However, because the processing speed increases as the number of templates in the pattern becomes larger, the problem of template management is usually encountered when using this method.

C. *Model-based classification*

The studies using a classification model-based approach is a way to model the sequence data and find the model most similar through matching to the new input. Mathew, et al. made a model using HMM to a sequence of visited locations to derive the destination [10]. Gambs, et al. determined the destination by using the mobility Markov chains from the sequence of the POI (Point of Interest) to make the model [11]. Burbey, et al. predicted the destination by applying the PPM using the residence time and visit time and location [12]. The studies using the model-based classification predict the destination by using only spatial information, such as GPS. However, when using the spatial information only, there is a problem that the prediction is biased to the lower information that lacks of initial movement.

III. THE PROPOSED METHOD

When it comes to represent the path using only the spatial information, prediction methods cannot resolve the problem of partially overlapped paths as shown in Fig. 2. This means that there are main trajectories for users to move. When the user passes through the main trajectory, using only spatial information will result in predicting an incorrect destination which overlaps the main trajectory. In order to work out this problem, we extend the path information by adding time and other context data. Location information, including temporal and spatial among others, is collected by sensors in smartphone environment. Time and date of the smartphone are used as temporal information, and latitude and longitude from the GPS sensor are utilized as spatial information. Also, other context data from accelerometer, magnetic, and orientation sensors determine the mode of transportation of the user.

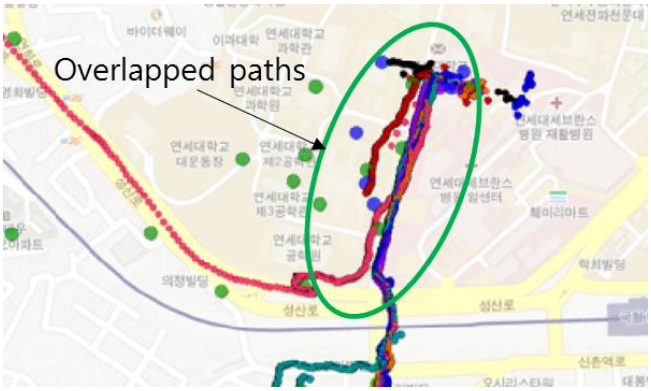

Figure 2. An example of overlapped paths

In this paper, in order to predict the destination, the path information is stored using hidden Markov model (HMM). HMM is a statistical model characterized by Markov process with unknown parameters, modeling observations to determine these hidden parameters. It is a widely used technique that stochastically models sequence data of the time series. It is mainly composed of the state transition probabilities, and the probabilities that select the observation value at each state.

 ISBN: 978-1-61208-366-7

About the path $P = \{L_1, L_2, \cdots, L_n\}$ of length n, the information of location L_t at time t is affected by the information of L_{t-1}. Therefore, it can be assumed to be a Markov process. In this method, because it uses a variety of information of context, we make a model using HMM, which probabilistically represents many features. A HMM is defined by state transition probability A, probability distribution of observed symbols B, and probability distribution of initial state Π. One HMM, λ, is expressed as (1).

$$\lambda = \{A, B, \Pi\} \quad (1)$$

HMM consists of only one probability distribution which is made from various sequence data. Therefore, it eliminates the storage of unnecessary sequence data because new input sequences can be compared with only one model, as opposed to comparing with many sequence data. Thus, HMM is suitable in mobile environment which has limitations in processing time and storage.

In this method, when m_i is the ith HMM that makes a model with the same paths as the departure and destination about $P_{input} = \{L_1, L_2, \cdots, L_t\}$ which is moving path to the time t, we find a HMM model $\hat{m}$ which is the most similar to P_{input}. Destination of $\hat{m}$ is predicted destination. Also after finding the $P_{optimal}$ which is the most similar to P_{input}, we pick up the future visiting locations based on the $P_{optimal}$. The structure of the proposed method is shown in Fig. 3.

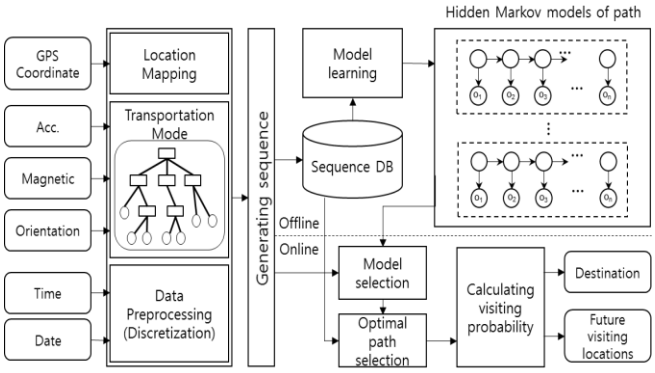

Figure 3. The structure of the proposed method

A. Construction of path sequences

The path sequence is a set of locations along the temporal flow. Each element of the sequence represents the place. The ith location of the path sequence P is represented by (2).

$$L_i = \{time_i, weekday_i, place_i, transportation_i\} \quad (2)$$

The L_i used for the observed symbol of HMM is quantized as follows.

The temporal information of the place, to generate the observed values, separated by time and day of the week. Because the user works in a different pattern on the basis of the time of day and day of the week, we extract the two features in the time information. The time, in the representation of the time zone, has a total of 6 values through the vector quantization, and the day of the week has 7 values.

Spatial information is formed by the latitude and longitude coordinates of a large number. However, as it is not possible to use up all coordinates, it is necessary to extract the key locations. The key locations may be the departure and destination locations or a location to visit during moving such as crossroad, bus stop, subway station, etc. In other words, a key location means user's meaningful location or intermediate location to it. Previous studies to extract key locations use K-means clustering which finds the centers of crowded area in GPS data [12].

However, this method should determine the number of locations for extracting key locations. Also, it cannot guarantee the performance of extraction, because the criteria of density are uncertain. Therefore, we used G-means clustering [13] for extracting key locations which follow a Gaussian distribution in GPS data [14]. The G-means clustering is a clustering method to test each cluster whether Gaussian distribution through statistical verification and repeat the K-means clustering until all clusters follow Gaussian distribution. Extracted locations were labeled by discrete value as the observation symbols for HMM.

Other context information is used by transportation in location. In order to determine the transportation mode, this method classifies the transportation based on accelerometer, magnetic, orientation sensors in smartphone. For converting high-level data to low-level data, this method uses decision tree algorithm which is suitable in mobile device because the recognition speed is faster than other methods. Transportation mode is classified into 4 states, such as Staying, Walking, Running and Vehicle [1].

TABLE II. QUANTIZATION OF INPUT DATA

Type	Low-level data	Quantized data
Temporal information	Time (0-23)	6 separate units for 4 hours
	Day of week	Day of week
Spatial information	GPS coordination	Labeled location
Other context information	Accelerometer sensor value	4 state (Staying, Walking, Running, Vehicle)
	Magnetic sensor value	
	Orientation sensor value	

B. Building path model based on HMM

A path which is a subsequence of trajectory can be generated from many different paths. Because it is not possible to create a model for all paths, we construct a HMM based on the paths of start and end points. The HMM has information about the start and end points of a path. In this way, the number of HMM is reduced, but is not enough because of the large number of locations. When the number of locations is n, $n \times (n-1)$ HMMs are required to create models. So, we build the models only with start and end points of a path for departure and destination.

The HMM with path information is learned by Baum-Welch algorithm [15], which is a learning method typically to

represent the probabilistic information of multiple sequences. Fig. 4 shows the generation process of HMM-based path model in a real data. In this figure, a HMM is created by a pair of departure and destination. This is an example to use the different sequences to model as the same path.

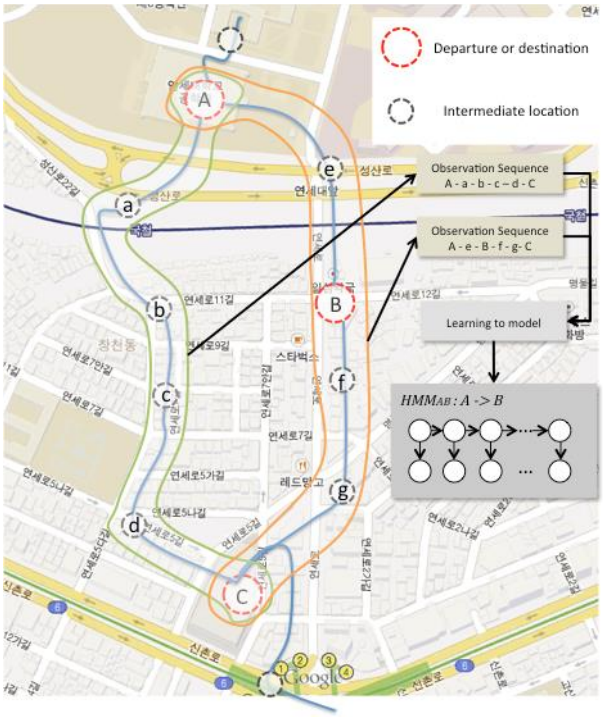

Figure 4. An example of building path model

C. Predicting destination

If a user reaches at the present location with path $P_{present}$ at time t, which is the same with the moved path of length T, P_{past} until time t in past, we assume that moving location of $P_{present}$, in the future, is the same with moving path of P_{past} on $t+1$ to T. Based on this, we define that destination of $P_{present}$ is a destination of HMM model which includes the most similar path with $P_{present}$ until time t.

The similarity of path can be represented by the probability that $P_{present}$ is observed until time t in HMM_{AB} whose departure is A and destination is B. When the sequence of states until time t is $Q = \{q_1, q_2, \cdots, q_t\}$ on λ_{AB} which is the model parameter of HMM_{AB}, the probability of observing $P_{present}$ can be expressed by (3).

$$\mathrm{P}(P_{present}|\lambda_{AB}) = \sum_Q \mathrm{P}(P_{present}, Q|\lambda_{AB})$$
$$= \sum_Q \mathrm{P}(P_{present}|Q, \lambda_{AB})\mathrm{P}(\mathrm{Q}|\lambda_{AB}) \quad (3)$$

$\mathrm{P}(P_{present}|Q, \lambda_{AB})$ is the probability to appear $P_{present}$ given sequence of state and model parameter of HMM_{AB}. Also, $\mathrm{P}(\mathrm{Q}|\lambda_{AB})$ is the probability to select Q given model parameter of HMM_{AB}. Therefore, the probability of observing $P_{present}$ in HMM_{AB} is the sum of $\mathrm{P}(P_{present}|Q, \lambda_{AB})\mathrm{P}(\mathrm{Q}|\lambda_{AB})$ for all the state sequences. Equation (4) is expressed by using Markov process on (3).

$$\sum_Q \mathrm{P}(P_{present}|Q, \lambda_{AB})\mathrm{P}(\mathrm{Q}|\lambda_{AB})$$
$$= \sum_Q (b_{q_1}L_1 \cdots b_{q_t}L_t)(\pi_{q_1} a_{q_1 q_2} \cdots a_{q_{t-1}q_t}) \quad (4)$$

$\mathrm{P}(P_{present}|\lambda_{AB})$ is calculated using (2). After calculating the probability of observing $P_{present}$ all about HMM, $HMM_{optimal}$ which has the highest probability is selected. Destination of $P_{present}$ is a destination of the selected $HMM_{optimal}$.

D. Calculating visit probabilities

Based on the destination which is determined by $HMM_{optimal}$, the probabilities of visiting desination and intermediate locations are calculated. First, we find out a path sequence $P_{optimal}$, which is the same with a departure and a destination of $HMM_{optimal}$ and includes the current path $P_{present}$. By the assumption of section C, the location movements of $P_{optimal}$ from time $t+1$ decide the future location movements of $P_{present}$.

Determining a sequence of future movements of the locations allows to find out optimal state sequence $\hat{Q}$ from $HMM_{optimal}$ about path $P_{optimal}$ and calculate the probabilities of visiting locations based on $\hat{Q}$. The method of calculating the optimal state sequence $\hat{Q}$ is as (5).

$$\hat{Q} = max_{Q=q_1 q_2 \cdots q_T} P(Q|P_{optimal}, \lambda_{optimal}) \quad (5)$$

This can be calculated using the Viterbi algorithm [16]. When the observation sequence O is given in the HMM, Viterbi algorithm searches for a state sequence Q as shown the best. That is, it is possible to find the state sequence $\hat{Q}$ that maximizes the probability of discovery of $P_{optimal}$ from the $HMM_{optimal}$.

If it finds the state sequence $\hat{Q}$ based on the Viterbi algorithm, when $HMM_{optimal}$ is given, the probability that path $P_{optimal}$=$\{L_1, L_2, \cdots, L_T\}$ and sequence of states $\hat{Q}$ are found together (joint-probability) is the same as (6).

$$\mathrm{P}(P_{optimal}, \hat{Q}|\lambda_{optimal})$$
$$= \prod_{t=1}^{T} P(\hat{Q}_t|\hat{Q}_{t-1}, \lambda_{optimal})P(L_t|\hat{Q}_t, \lambda_{optimal}) \quad (6)$$

Based on (6), when the user has actually moved to the location of the ith path $P_{optimal}$ whose length is T, the probability of the location L_j of the jth $(j > i)$ may be calculated by the following equation (7).

$$\mathrm{P}(L_j) = \prod_{k=i}^{j-1} P(\hat{Q}_{k+1}|\hat{Q}_k)P(L_{k+1}|\hat{Q}_{k+1}) \quad (7)$$

It is possible to calculate the probabilities of the intermediate locations and a destination previously visited using (7).

 ISBN: 978-1-61208-366-7

IV. EXPERIMENTS

The experiments were performed with the sensor data that 10 university students in their 20s collected during the five months with the smartphone (Samsung Electronics, smartphone SHV-E300K). The description of each user's data is shown in Table III.

TABLE III. DESCRIPTION OF DATA

	#Location	#Path	Size of storage
User 1	16	193	2.44GB
User 2	20	268	2.62GB
User 3	32	149	1.46GB
User 4	50	288	3.41GB
User 5	42	309	3.66GB
User 6	32	233	1.34GB
User 7	28	236	1.48GB
User 8	24	294	3.21GB
User 9	36	237	2.37GB
User 10	14	189	2.08GB

TABLE IV. ACCURACY ACCORDING TO ADVANCEMENT OF THE PATH

	0% (Departure)	20%	40%	60%	80%	100%
User 1	48.05	62.34	75.32	87.01	88.31	90.91
User 2	51.11	60.00	73.33	91.11	93.33	97.78
User 3	50.00	58.06	74.19	87.1	93.55	95.16
User 4	38.46	58.46	67.69	78.46	87.69	89.23
User 5	84.62	87.18	89.74	94.87	94.97	94.87
User 6	74.21	77.24	81.75	85.32	87.86	93.38
User 7	68.90	74.71	78.89	82.36	86.48	88.11
User 8	62.20	67.89	74.71	79.54	85.61	90.71
User 9	53.90	56.74	59.24	65.96	70.71	75.67
User 10	48.12	52.87	68.12	78.22	83.47	93.11
Average	57.96	65.55	74.30	83.00	87.20	90.89

In order to evaluate the accuracy of the proposed detination prediction method, we measured the accuracy according to the progress of path. Prediction result of the advancement of the path is illustrated in Table IV. Looking at the prediction accuracy in accordance with the progress of the path, as the path is largely moves, it can be seen that the prediction accuracy becomes higher because the information of the location movement is increased. 0% progression of the path, that is, is capable of predicting only location information from the starting location. HMM showed accuracy of 57.96% on average only with the information of departure.

Fig. 5 shows the accuracies for the data of 10, indicating the average of the predicted test results with and without the use of context information and progress of the path. When using all context information, the accuracy is the the highest. When using only the spatial information, the accuracy is the lowest. The difference in the case of not using the transportation information and time information is not large, and by using the information of the day, it can be seen that the accuracy is significantly increased.

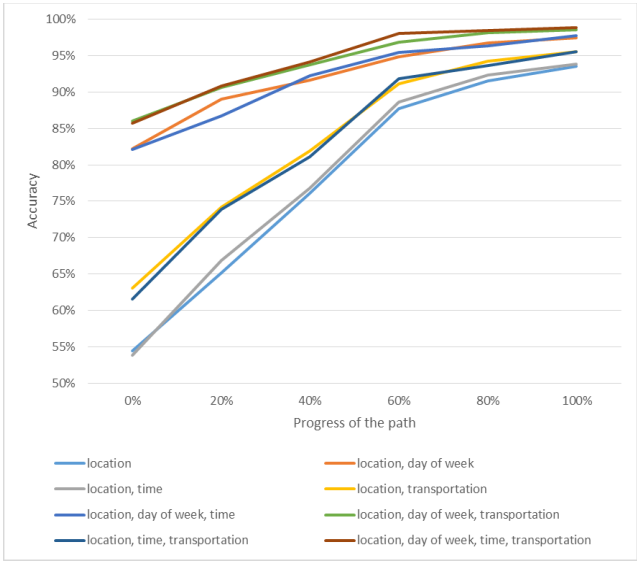

Figure 5. Accuracy according to context contribution and advancement of the path

To demonstrate the usefulness of HMM for the prediction of the destination, we compared it with dynamic time warping (DTW) method [8] which is a template pattern matching method. Fig. 6 shows the average accuracy of prediction based on the data of the 10 users. When the path of the progress is 0% in DTW, because of the shorter length of the input, the prediction is impossible. In the case of HMM, $P_{present}$ matches up the part of the path of the past. However, in DTW, because it matches the full path of the past, it is shown in very low prediction probabilities.

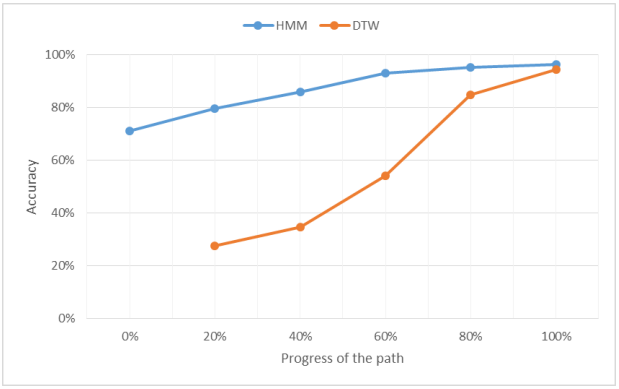

Figure 6. Comparison of accuracy of HMM and DTW

 ISBN: 978-1-61208-366-7

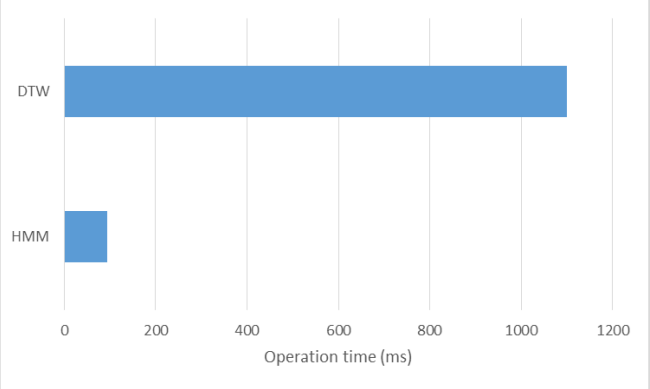

Figure 7. Comparison of processing time of HMM and DTW

Fig. 7 illustrates the prediction time of DTW and HMM. It shows the overall average of time that has been used for predicting each input path. It is possible to see the HMM 12.67 times faster than DTW. In the DTW, the amount of calculation per one template is bigger than HMM because of matching the sequence of the entire path in many cases. Also, it consumes a lot of time, since input data are matched to all the patterns which are the same as departure and destination. However, in the HMM, because it calculates the similarity only until current moving time and it makes one model for all the paths which have the same departure and destination, its running time is shorter.

V. CONCLUSIONS

In this paper, we have proposed a destination and intermediate location prediction method using user's smartphone sensor data. A path is the changes in the location due to human judgment. Based on this, we represent the path model using the HMM where the user moves, and predict a destination. The pre-processing for destination prediction includes extracting key locations, and classifying transportation mode using smartphone sensor data. After making a HMM of paths using pre-processing data, HMM is to learn the parameters. When new input comes, this method finds out the optimal HMM and decides a destination. Also, it calculates the probabilities of visiting destination and intermediate locations. When evaluated with the data of 10 users' destinations, by using not only the spatial information, but a variety of context information improves the accuracy significantly. Also, when compared to the other methods, this method yielded higher accuracy and showed fast running time.

ACKNOWLEDGENMENTS

This work was supported by Samsung Electronics, Inc.

REFERENCES

[1] Y. J. Kim, and S.-B. Cho, "A HMM-based location prediction framework with location recognizer combining k-nearest neighbor and multiple decision trees," 8th Int. Conf. on Hybrid Artificial Intelligent Systems, pp. 618-628, 2013.

[2] I. Burbey and T. L. Martin, "Predicting future locations using prediction by partial match," First ACM Int. Workshop on Mobile Entity Localization and Tracking in GPS-less Environments, pp. 1-6 , 2008.

[3] Z. Xing, J. Pei, and E. Keogh, "A brief survey on sequence classification," ACM SIGKDD Explorations Newsletter, vol. 12, no. 1, pp. 40-48, 2010 .

[4]] T. M. T. Do, and D. Gatica-Perez, "Where and what: Using smartphones to predict next locations and applications in daily life," Pervasive and Mobile Computing, pp. 79-91, 2013.

[5] Z. Lu, Y. Zhu, V.W. Zheng, and Q. Yang, "Next place prediction by learning with multiple models," Mobile Data Challenge Workshop, 2012.

[6] A. Monreale, F. Pinelli, and R. Trasarti, "WhereNext: a location predictor on trajectory pattern mining," 15th Int. Conf. on Knowledge Discovery and Data Mining, pp. 637-646, 2009.

[7] S. Akoush, and A. Sameh, "Mobile user movement prediction using Bayesian learning for neural networks," 2nd Int. Conf. on Systems and Networks Communications, pp. 191-196, 2007.

[8] B. Kim, J. Y. Ha, S. Lee, S. Kang, and Y. Lee, "AdNext: A Visit-Pattern-Aware mobile advertising system for urban commercial complexes," 12th Workshop on Mobile Computing Systems and Applications, pp. 7-12, 2011.

[9] S.-H. Lee, and B.-K. Kim, "A path prediction method using previous moving path and context data," Int. Symposium on Advanced Intelligent Systems, pp. 199-202, 2009.

[10] W. Mathew, R. Raposo, and B. Martins, "Predicting future locations with hidden Markov models," ACM Conf. on Ubiquitous Computing, pp. 911-918, 2012.

[11] S. Gambs, M. O. Killijian, and M. N. del Prado Cortez, "Next place prediction using mobility Markov chains," First Workshop on Measurement, Privacy, and Mobility, p. 3, 2012.

[12] A. J. Dou, V. Kalogeraki, D. Gunopulos, T. Mielikinen, V. Tuulos, S. Foley, and C. Yu, "Data clustering on a network of mobile smartphones," IEEE/IPSJ Symposium on Applications and the Internet , pp. 118-127, 2011.

[13] G. Hamerly, and C. Elkan, "Learning the k in k means," Advances in Neural Information Processing Systems, vol. 16, pp. 281, 2004.

[14] M. Kim, D. Kotz, and S. Kim, "Extracting a mobility model from real user traces," Int. Conf. on Computer Communications, vol. 6, pp. 1-13, 2006.

[15] L. E. Baum, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," Ann. Math. Statist, vol. 41, pp.164 – 171, 1970.

[16] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, vol. 77, pp. 257–286, 1989.

 ISBN: 978-1-61208-366-7