2017 IEEE 18th International Conference on Mobile Data Management

UDAT: User Discrimination using Activity-Time Information

Snigdha Das*, Dibya Jyoti Roy*, Subrata Nandi†, Sandip Chakraborty*, Bivas Mitra*
*Dept. of CSE, IIT Kharagpur, India; †Dept. of CSE, NIT Durgapur, India
Email: {snigdha00582,djr.kgp,subrata.nandi}@gmail.com, {sandipc,bivas}@cse.iitkgp.ernet.in

***Abstract*—This paper explores the feasibility of automatically discriminating users from the activity as well as temporal information of their daily routine. We observe that everyone pursues a daily semi-regular activity pattern. Based on this observation, we have developed a system *UDAT* and experimented on Microsoft Geolife as well as UDAT datasets. With Geolife transportation activity log and UDAT motion-static activity log, the system achieves 73.3% and 80.68% accuracy, respectively. Although the overall system accuracy is moderate, the system achieves the highest accuracy when the users belong to the different activity buckets. This signifies the utility of two-phase classification for user discrimination.**

***Keywords*-user identification; activity based classification;**

I. Introduction

Increasing availability of sensor equipped smartphone provides new opportunities for development of various context-aware services [1]. In this line, activity recognition [2] is coming up as a mainstream research agenda in the domain of smartphone computing. Among the existing activity recognition techniques, wearable devices [3] and smartphone [4] based data tracking schemes have become most popular due to the minimal user involvement. This is imperative to realize the fact that activity pattern, followed by an individual, carries a signature of that person. For instance, monitoring Bob's daily routine in the weekdays, one can notice that Bob goes out for jogging at 6.00 and leaves for his office at around 8.00 via subway and returns to home at around 16.00. On the other side, Alice walks into her school around 9.00 and return to home nearly at 16.00. In this example, this is clearly visible that the daily activity patterns of Alice and Bob carry their signature, and if leveraged properly, they may recognize (at least demarcate) the individual persons.

Automatic identification and discrimination of users may work as a core for different futuristic applications. For instance, in the domain of "Smart City," personalized recommendation is inevitable for the development of the next generation services. In smart-office, different smart devices like smart lights, smart door-locking systems, smart room temperature control should automatically recognize individual residents and take personalized actions according to her requirement. In this context, each time registering individual users to the system may appear restrictive; the system should automatically identify the users in a privacy preserved manner without accessing any personally identifiable information (like International Mobile Equipment Identity, etc.) and perform service differentiation. In [4], Kwapisz et al. proposed a user identification model from the accelerometer data [5]. Murmuria et al. [6] studied the user's physical movement pattern, touch pattern, and power usage to discover the users' unique characteristics. Importantly, most of the endeavors above consider only the short period data trace, even in a controlled environment. However, in a realistic setup, and in a long time span, the performance of these models may get severally compromised. Few attempts [7], [8] have been made in bits and pieces in discovering the daily activity patterns through human activity monitoring. Huynh *et al.* [9] incorporated the indoor as well as outdoor activities of a user to develop the daily routine recognition model. Considering the limitations of the state of the art user identification methodologies and observing the progress made in the gamut of activity recognition, there exists a scope to explore the potential of the daily activity patterns of user identification (and discrimination).

The objective of this paper is to develop $UDAT$, a user identifier cum discriminator model leveraging on the regularities present in the user activity pattern. The model is expected to differentiate two users only based on the collected trace of activities. First, we develop $UDAT$, a user discrimination model, which involves three major modules – (a) Activity based classification, (b) Outliers detections, and (c) Temporal classification (§ III). We differentiate the users only based on the dominant activities and then construct one class for each activity and populate the classes with corresponding users. In the second step, we rely on the temporal activity signals (say starting time, duration, etc.) and classify the users in each activity class. We introduce two real datasets (a) Microsoft Geolife dataset (b) UDAT dataset to evaluate the performance of the model (§ II). We show that $UDAT$ model discriminates users with 73.3% and 80.68% accuracy, for Microsoft Geolife and $UDAT$ datasets respectively and outperforms the baseline algorithms (§ IV and § V).

II. Dataset

In this paper, we develop and experiment our model on the following two real datasets (a) Microsoft Geolife dataset [10], (b) UDAT dataset that is collected in-house. Both the datasets consist of a set of users $x_i \in X$ performing a set of activities $a_j \in \mathcal{A}$. For each user x_i, it stores an information vector (t^i_j, a^i_j, s^i_j) which describes that the user x_i performs the activity a_j at time stamp t_j. Furthermore, it records the raw

2375-0324/17 $31.00 © 2017 IEEE
DOI 10.1109/MDM.2017.59

sensor data $s_j \in \mathcal{S}$. s_j^i represents the sensor reading associated with activity a_j at time t_j for the user x_i.

A. Microsoft Geolife Dataset

Microsoft Geolife dataset [10] contains GPS trajectories of 178 users in a period of four years. The trajectory data is logged by different GPS loggers and GPS-enabled smartphones. The dataset describes latitude, longitude, altitude information along with the time-stamp sequence for all the users. Out of all 178 users, 24 users' data contains activity information as ground truth. The activities depict the various transportation modes that the user have chosen during the data recording period. The transportation modes considered in the dataset are 'Bike,' 'Bus,' 'Car,' 'Subway,' and 'Walk.'

B. UDAT Dataset

We develop an Android data collection application running on smartphones, which seamlessly collect the accelerometer data along with the activity label for each user. The application works on top of the Google Activity Recognition API for labeling the user activities along with a confidence value. This API generates 8 activity labels such as InVehicle, OnFoot, OnBicycle, Running, Walking, Still, Tilting, and Unknown. Out of these 8 activities, only Tilting can be performed together with some other activities. We have used Earth gravity for removing the conflicts generated due to mobile devices orientation in the accelerometer data. We have recruited total 15 volunteers (undergraduate and postgraduate students) with age group 20 - 35 (two females) and conducted the data collection experiments for three months. The subjects are instructed to carry their smartphones with them for all the time. The devices automatically track user activities and sensor data using Google activity recognition API and sensor listeners, respectively. The collected data are temporarily stored within the smartphones. Subsequently, the data gets uploaded periodically to the central server.

III. Developing *UDAT* Model

In this section, we propose *UDAT*, a user discrimination model leveraging on the regularities in user activity pattern. The core of the UDAT model (see Fig. 1) composed of three major modules – (a) Activity based classification, (b) Outliers detections, and (c) Temporal classification. In UDAT, we discriminate users using two modules – Activity-based classification and Time-based classification. Outlier detection module is responsible for mining the normal activity pattern of the users and eliminating the outlier data points.

A. Activity based Classification

The objective of this module is to classify users only based on the activities performed by them. We construct one bucket A_a for each activity $a \in \mathcal{A}$ and populate the activity buckets with the users performing the corresponding activities. We consider the *day wise* activity log data of each user x_i and identify his two major activities p and q based on the total time spent on those activities. We insert all the timestamps t_p^x and

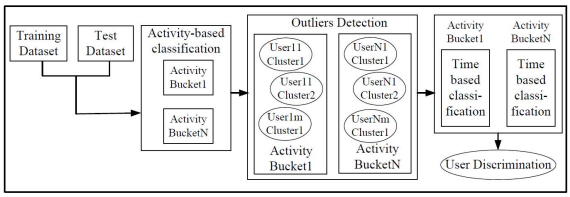

Fig. 1. User Discrimination Model

t_q^x, the occurrence of the activities p and q by user x_i, in the activity buckets A_p and A_q respectively. Hence, each activity bucket contains information (timestamps) of several users. The intuition behind the module is the following. Consider three users $U1$, $U2$ and $U3$ where activity p is dominant for users $U1$ and $U2$ whereas activity p is dominant for users $U2$ and $U3$. Hence users $U1$, $U2$ fall into activity bucket A_p and $U2$, $U3$ fall into bucket A_q. As a result, $U1$ and $U3$ can be easily differentiated by simple checking the activity buckets.

In Geolife dataset, we observe 10 activity buckets (corresponding to respective transportation modes such as airplane, bike, boat, bus, car, running, subway, taxi, train, and walking) whereas in UDAT dataset, we obtain 5 activity buckets (representing activities such as InVehicle, OnBicycle, OnFoot, Still, and Unknown). Running and Titling are not listed in the bucket as the users are not performing these activities frequently.

B. Outliers Detection

A close monitoring of user x_i in activity bucket A_p reveals that, timestamps t_p^x captures normal activity (p) patterns of x_i, mixed with few exceptions. In this module, we wish to eliminate those exception patterns through outlier elimination. We consider the following features to characterize the normal & outlier activity patterns (a) first start time of the activity, and (b) activity trip count. For user x_i, we compute the feature vector of all the days and apply *DBSCAN* model with epsilon (Eps) value 1 and minimum sample size 15. The outliers detection model identifies the users' multiple patterns of activity. We observe that for our datasets at most two patterns exist per activity for a user. Incidentally, we observe few users without any proper cluster formation; we eliminate those users (having no normal activity pattern) from the respective activity bucket. In Geolife dataset, after outlier elimination, we obtain only 5 activity buckets out of total 10 buckets whereas, in UDAT dataset, we obtain only 4 activity buckets out of total 5 buckets; this results due to the elimination of all the users from those respective buckets. Finally, after outlier elimination, we obtain six valid users for both the datasets.

C. Temporal Classification

Finally, in this module, time information based classification is performed for discriminating the users inside each activity bucket. We rely on the following four temporal features (computed per day) for classifying users (a) first start time of the activity, (b) total duration for that activity, (c) maximum activity duration, and (d) activity trip count. We compute

these features for each day and construct the feature vector considering all the days in the dataset. We implement several classification techniques such as k-NN, Random Forest, Logistic Regression, and Support Vector Machine for user discrimination inside each activity bucket.

IV. Experimental Setup

The experimentation procedure of $UDAT$ model is discussed in this section. Both for Geolife and UDAT datasets, we feed the time-stamped raw activity sequence to the $UDAT$ model. The model then applies activity based classification from both the datasets. During outliers detection, normal points are separated, and we receive six users each out of 24 and 15 users in Geolife and UDAT dataset, respectively. The filtered normal data points are used for time-based classification. We use $2:1$ ratio of training and testing datasets for this final classification. All the feature vectors are annotated with user ID for learning and validating our model. We execute this classification procedure for 25 times for removing the selection bias of the training set. Accuracy is used for assessing the overall system performance. We evaluate the performance of $UDAT$ from two different perspectives. First, we implement a competing algorithm, proposed in [4] and evaluate it on our dataset. Next, we generate two variations of $UDAT$ as baselines, called $UDAT\_OnePhase$ and $UDAT\_Outlier$, and compare them with $UDAT$. Essentially, these baseline models demonstrate the justification of the individual modules of $UDAT$.

A. Competing Model

In [4], Kwapisz *et al.* leveraged on the accelerometer data for user discrimination. The model recorded the accelerometer data with duration of 10 seconds from the users' mobile devices for all the activities such as walking, jogging, and climbing in a controlled environment. The captured data were used for generating the features such as mean, standard deviation, and absolute difference. The generated features along with the user ID were used for classifying the users. In UDAT dataset, we compute the mean and standard deviation of all axises of accelerometer value. The geolife dataset does not contain accelerometer data log; hence we replace it with the GPS co-ordinates (latitude and longitude) as the sensor signal and use this information to generate the feature set. However, for both the datasets, the principle of their model remains invariant.

B. UDAT_OnePhase

In this baseline model, we propose a simple variation of $UDAT$ where we classify the users directly based on the temporal features, such as first start time of the first performed activity, total duration for all the activities, maximum trip duration, and trip count. Essentially, in this model, we only use the 'Time-based classification' module from $UDAT$.

TABLE I
Geolife Dataset: Accuracy (%) comparison of different models (**A**: Activity, **CM**: Competing Model)

A	*UDAT*	*CM* [4]	A	*UDAT*	*CM* [4]
Bike	85.47	69.47	Bus	82.92	65.03
Car	83.20	79.46	Subway	70.50	68.68
Walk	44.41	32.44	**Average**	73.30	63.01

C. UDAT_Outlier

We use another variation of $UDAT$ for comparison, by dropping the outlier detection module from the model. Hence this baseline model is simply a combination of the 'activity based classification' and 'temporal classification' modules. The feature set remains identical to the $UDAT$ model. Comparison with this baseline shows the advantage of the outlier detection module in our model.

V. Experimental Results

In this section, we evaluate the performance of $UDAT$ against the competing and baseline methods given in [4] and the two variants of the $UDAT$ model.

A. Overall System Performance

Geolife dataset – We compare the $UDAT$ system accuracy with the $Competing$ model for both Geolife and UDAT datasets. Table I shows the accuracy across different activities in Geolife dataset. For $UDAT$ model, individual activity buckets archive more system accuracy than the *competing* model. Finally, we receive overall system accuracy for the $UDAT$ model as 73.3% whereas this value is 63% for $Competing$ model. The $UDAT\_OnePhase$ model also obtains very low accuracy 20% for this dataset.

UDAT dataset – We demonstrate the performance of the proposed model on the UDAT dataset against the $Competing$ algorithm given in [4] and the baseline $UDAT\_Outlier$ model. The accuracy comparison chart for UDAT dataset is shown in Table II. We observe that $UDAT$ outperforms for all the activity cases. Although a few activity buckets excluded due to the outlier elimination mechanism in the $UDAT$ model, the rest of the activities such as OnBicycle, OnFoot, Still, and Unknown obtain high accuracy 100%, 65.07%, 84.64%, and 73.01%, respectively. As the outlier detection in $UDAT$ model eliminates data point with no pattern, we are receiving a few activity buckets (say Tilting) with no user (marked as NA). The overall accuracy obtained from $UDAT$, $Competing$ and $UDAT\_Outlier$ models are 80.68%, 34.32%, and 40.35%, respectively. Additionally, the $UDAT\_OnePhase$ model also provides low accuracy (20%).

B. Implication of Different Machine Learning Algorithms

We experiment time-based classification module of UDAT model using different machine learning techniques. The comparative study of the accuracy values is shown in Fig. 2. All the techniques achieve more than 70% accuracy for UDAT dataset. However, RandomForest surpasses the other models such as kNN, Logistic Regression, and SVC.

TABLE II
UDAT DATASET: ACCURACY (%) COMPARISON OF DIFFERENT MODELS

Activity	*UDAT*	*Competing* [4]	*UDAT_Outlier*
Tilting	NA	39.29	31.13
InVehicle	NA	32.65	33.82
OnBicycle	100.00	22.36	48.03
OnFoot	65.07	32.92	40.78
Running	NA	32.77	NA
Still	84.64	33.16	46.75
Unknown	73.01	24.56	41.61
Walking	NA	56.86	NA
Average	**80.68**	**34.32**	**40.35**

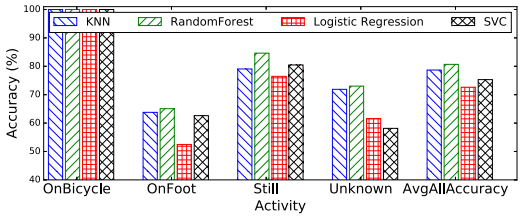

Fig. 2. Accuracy chart of UDAT dataset using different classifier

TABLE III
UDAT DATA RECALL

Users	OnFoot	Still	Unknown	Avg Recall
KGP_U1	NA	58.04	NA	58.04
KGP_U2	NA	75.85	62.14	69.00
KGP_U3	65.41	89.61	NA	77.51
KGP_U4	NA	96.11	NA	96.11
KGP_U5	64.06	86.50	78.65	76.40
KGP_U6	NA	100.00	NA	100.00

TABLE IV
UDAT DATA PRECISION

Users	OnFoot	Still	Unknown	Avg Precision
KGP_U1	NA	62.41	NA	62.41
KGP_U2	NA	75.21	60.42	67.81
KGP_U3	69.20	85.03	NA	77.12
KGP_U4	NA	100.00	NA	100.00
KGP_U5	59.98	89.10	79.85	76.31
KGP_U6	NA	93.65	NA	93.65

C. User Specific System Performance

The user-wise recall and precision values of the $UDAT$ model for UDAT dataset are shown in Table III and Table IV, respectively. Furthermore, average recall of the activity bucket also represents the average accuracy of the users. Few cells contain *NA* as there exists no pattern in activity bucket after the outlier detection of $UDAT$ for that user. We observe that user KGP_U6 is most likely to be identified in the dataset as the average accuracy is 100% whereas user KGP_U1 is least likely to be recognized among the six users due to the activity overlap with other users. Precision value also depicts almost the similar phenomenon.

D. 'Per Day' Activity Sampling in 'Activity Classification'

In the 'Activity classification module' of section III, we sample the *day wise* activity log data of each user for populating the activity buckets. In this section, we attempt two variations, implementing the sampling based on (a) full day and (b) half day activity logs of all the users. Although each activity bucket contains similar users; the feature set is mostly violated by the 12 and 24 hours patterns. The detailed results are shown in Table V. We obtain lower accuracy for half day value based classification. Therefore, we conclude that the users' routine is more regular for the full day rather than the half day.

TABLE V
FULL DAY VS HALF DAY ACCURACY (%) MEASURE (**FD**: FULL DAY SAMPLING, **HD**: HALF DAY SAMPLING)

Activity	FD	HD	Activity	FD	HD
Bike	85.47	45.30	Bus	82.92	42.63
Car	83.20	63.05	Subway	70.50	33.26
Walk	44.41	25.79	**Average**	73.30	42.01

VI. CONCLUSION

The major contribution of this paper is to demonstrate the signatures embedded within the daily activity patterns as a valid alternative to the conventional sensor-driven user identification paradigm. As a proof of concept, we have developed $UDAT$, a user identifier cum discriminator model, leveraging on the regularities present in the user activity pattern. In this model, we have leveraged on two key modalities – activity differentiation and temporal variations for discriminating the users. Experimental results on Microsoft Geolife and UDAT dataset demonstrated the promising results with overall system accuracy 73.3% and 80.68%, respectively; it outperformed the sensor-driven competing algorithms in both the cases.

REFERENCES

[1] Ö. Yürür, C. H. Liu, Z. Sheng, V. C. Leung, W. Moreno, and K. K. Leung, "Context-awareness for mobile sensing: a survey and future directions," *IEEE Communications Surveys & Tutorials*, 2014, vol. 18, no. 1, pp. 68–93.

[2] J. W. Lockhart, T. Pulickal, and G. M. Weiss, "Applications of mobile activity recognition," *ACM Ubiquitous computing*, 2012, pp. 1054–1058.

[3] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Communications Surveys & Tutorials*, 2013, vol. 15, no. 3, pp. 1192–1209.

[4] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Cell phone-based biometric identification," *IEEE Biometrics: Theory Applications & Systems*, 2010, pp. 1–7.

[5] M. Wolff, "Behavioral biometric identification on mobile devices," *International Conference on Augmented Cognition*, 2013, pp. 783–791.

[6] R. Murmuria, A. Stavrou, D. Barbará, and D. Fleck, "Continuous authentication on mobile devices using power consumption, touch gestures and physical movement of users," *International Workshop on Recent Advances in Intrusion Detection*, 2015, pp. 405–424.

[7] D. J. Cook, N. C. Krishnan, and P. Rashidi, "Activity discovery and activity recognition: A new partnership," *IEEE Transactions on Cybernetics*, 2013, vol. 43, no. 3, pp. 820–828.

[8] J.-H. Chiang, P.-C. Yang, and H. Tu, "Pattern analysis in daily physical activity data for personal health management," *Pervasive and Mobile Computing*, 2014, vol. 13, pp. 13–25.

[9] T. Huynh, M. Fritz, and B. Schiele, "Discovery of activity patterns using topic models," *ACM Ubiquitous computing*, 2008, pp. 10–19.

[10] Y. Zheng, L. Wang, R. Zhang, X. Xie, and W.-Y. Ma, "Geolife: Managing and understanding your past life over maps," *IEEE Mobile Data Management*, 2008, pp. 211–212.