GPS Trajectory Biometrics: From Where You Were to How You Move

Sami Sieranoja(✉), Tomi Kinnunen, and Pasi Fränti

School of Computing, University of Eastern Finland, P.O. Box 111, FIN-80101 Joensuu, Finland
{samisi,tkinnu,franti}@cs.uef.fi

Abstract. In this paper, we study how well GPS data can be used for biometric identification. Previous work has considered only the location and the entire route trajectory pattern. These can reveal the user identity when he repeats his every day moving patterns but not when traveling to new location where no route history is recorded for him. Instead of the absolute location, we model location-independent micro movements measured by speed and direction changes. The resulting short-term trajectory dynamics are modelled by Gaussian mixture model - universal background model (GMM-UBM) classifier from speed and direction change features. The results show that we can indentify users from OpenstreetMap data with an equal error rate (EER) of 19.6 %. Although this is too modest result for user authentication, it indicates that GPS traces do contain identifying cues, which could potentially be used in forensic applications.

1 Introduction

Thanks to smart devices combined with an increasing number of social media applications, collecting and sharing of personal data has never been easier as it is today. Besides photo and video uploads, smart-phones provide direct or processed information of the user's location or behavior via *global positioning system* (GPS), accelerometer or other sensor data. As an example, a sportswoman might upload her running route coordinates along with physical performance data.

GPS coordinate data contains a rich source of information about the user's whereabouts and behavior. This information can be used to provide useful services such as recommending potential friends based on user's trajectories [17]. On the other hand, it also raises a question of privacy [4].

Location-related or *spatial* cues include the most commonly used locations (such as user's home) or routes (such as daily route from home to work). Once combined with *temporal* (time-stamp) information, one is able to, for example, infer the future movements of a user [14] and the most likely times she will be absent from her home [3]. Speed estimates can be used for inferring the most likely means of transport (walking, bicycling, driving) [15] or whether the user respects speed limits.

A. Robles-Kelly et al. (Eds.): S+SSPR 2016, LNCS 10029, pp. 450–460, 2016.
DOI: 10.1007/978-3-319-49055-7_40

Even if a GPS trajectory data would be anonymized by obscuring the obvious identifying information, such as name and home address, the user might still be *re-identifiable* by linking an anonymized GPS coordinate data with non-anonymized data in the user's public profile in a social media application. Such information could be very sensitive; examples might be visit to an abortion clinic, church or premises of a political party [10].

In this study we focus on *user identification* based on GPS trajectory data. Differently from prior work that use user's location history for identification (*where you were*) [10,13], we approach the problem as a biometric identification task: to identify the user based on his or her physical or behavioral characteristics, but independently of the location or absolute timing of the trajectory data. We view the GPS trajectory coordinates of a person as an inaccurate measurement of the physical behavior of the user related to his or her muscle activity, such as gait or the way of steering a bicycle.

Our primary goal is to find out whether and how much of person-identifying traits exists in GPS trajectory data. Unlike [10,13] where the question was approached by the possibility to identify users by only identifying individual routes, we approach it as a statistical pattern recognition problem. That is, we model the distribution of short-term feature vectors derived from a set of GPS routes that reflect user's physical activity, rather than the locations visited. This way we are able to obtain a more accurate picture of how well users could be identified in situations where the training and test routes originate from different locations or dates. We utilize two public datasets to study the question whether person identification is feasible from GPS trajectory data, and if so, how much training and testing data is required.

2 Related Work

2.1 Location Privacy

The topic of location privacy has been a subject of many studies [6]. Of recent work, route uniqueness has been studied in [10] where it was shown that even low resolution mobile traces collected from mobile phone carriers are highly unique. Selecting only three points of a trace was enough to uniquely identify most traces. Similarly in [13] it was shown for GPS data that even when points are sampled out of the routes, they can still be reliably linked with the original routes.

2.2 Spatio-Temporal Similarity

Considerable amount of work has been devoted on *recommendation* based on GPS trajectory. As an example, in [17], potential friends are recommended based on user's trajectories. So-called *stay cells* are created based on detected stops. They are considered important since the user stayed there longer time. Similarity of trajectories is then measured based on *longest common subsequence* (LCS)

and giving more importance to longer patterns. In [8], revised version of LCS is applied by partitioning the trajectories based on speed and detected turn points. A similarity score is computed using both geographic and semantic similarity.

In [1], similarity of a person's days is assessed based on the trajectory by discovering their semantic meaning. The data is collected from tracking users' cars and pre-processed by detecting stop points. Most common pairs of stops are assumed to be user's home and work locations. Dynamic time warping of the raw trajectories using geographic distances of the points was reported to work best. In [16], personalized search for similar trajectories is performed by taking into account user preferences of which parts of the query trajectory is more important.

Complete trajectories are not always available and the similarity must then be measured based on sparse location data such as visits, favorite places or check-ins. In [7], user data is hierarchically clustered into geographic regions. A graph is constructed from the clustered locations so that a node is a region user has visited, and an edge between two nodes represents the order of the visits to these regions. This method still relies on the order of the locations visited.

Algorithm 1. Features for Segment

Input: Route segment S, Window width w
Output: Set of feature vectors F
  procedure FeaturesForSegment(S,w)
    $F \leftarrow \emptyset$
    for all $s_i \in S$ **do**
      $W \leftarrow (s_i, \ldots, s_{i+w})$
      speed $\leftarrow$ speed(W,w)
      turns $\leftarrow$ turns(W,w)
      $F_{\text{speed}} \leftarrow$ DftFeature(speed,w)
      $F_{\text{turns}} \leftarrow$ DftFeature(turns,w)
      $F_i \leftarrow (F_{\text{speed}}, F_{\text{turns}})$
    end for
  end procedure

Algorithm 2. DFT Feature

Input: Local window W, Window width w
Output: Feature X
  procedure DftFeature(W, w)
    $X \leftarrow (W - \text{mean}(W)) \circ \text{HammingWindow}(w)$
    $X \leftarrow |\text{fft}(X)|$
    $X \leftarrow \text{dct}(\log10(X))$
    $X \leftarrow X(1:24)$
  end procedure

3 Statistical User Characterization Using Short-Term GPS Dynamics

We model GPS user behavior by first calculating *discrete Fourier transform* (DFT) features from local speed and direction changes (Sect. 3.1). Then Gaussian mixture model - universal background model (GMM-UBM) classifier is trained on these features (Sect. 3.2).

3.1 Short-Term GPS Dynamics

DFT features are calculated from speed and turn angle data (see Algorithms 1 and 2). The route is processed using a sliding window of 100 s (100 or 50 points, depending on sample rate). Speed and turn angles are then calculated for each point inside the window. Turn angles are further processed by integration to produce a turn angle measure similar to what speed is to acceleration.

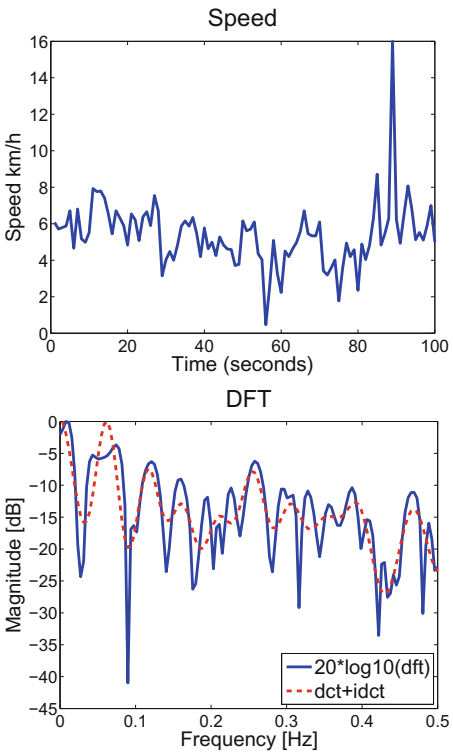

Fig. 1. DFT feature processing [user1]

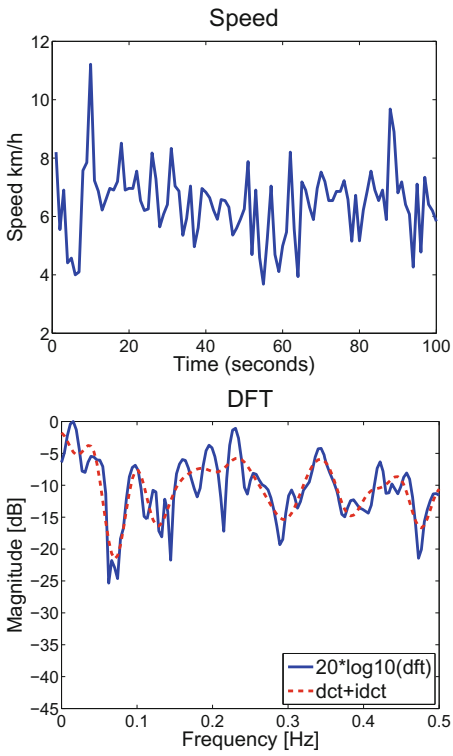

Fig. 2. DFT feature processing [user2]

After that, DFT is calculated separately for speed and turn angle. For each type of feature and each short-term segment, this yields complex-valued spectrum represented in polar form as $X(\omega_k) = |X(\omega_k)|e^{i\theta(\omega_k)}$, where $|X(\omega_k)|$ and

$\theta(\omega_k)$ denote, respectively, the magnitude and phase of the k^{th} frequency component. Similar to short-term speech processing, we discard the phase part. Logarithm of the retained part, magnitude, is then parameterized using discrete cosine transform (DCT) for dimensionality reduction and decorrelation purposes. We retain the first 24 DCT coefficients (including the DC coefficient) and concatenate the speed and turn angle features to yield feature vectors of dimensionality $24 \cdot 2 = 48$.

The DFT feature was designed to model how frequently and by how much the speed and turn angles change in a route segment. Additionally, they also reflect the speed of the user to some extent. The greater the speed, the more frequently there are turns and deacceleration/acceleration. This is expected to shift the part of the spectrum that correlates with road network to the right (to higher frequencies).

Figures 1 and 2 illustrate the features for two different users. In this example, user2 has less low frequency variations in the speed of the segment and this shows as a dip between frequencies 0.05 Hz and 0.1 Hz whereas user1 has a spike in same frequency range. Additionally, "dct + idct" represents a reconstruction of the original magnitude spectrum where inverse DCT is applied on the zero-padded feature vector containing the lowest 24 coefficients. The prominent characteristics of the magnitude spectrum are reasonably well preserved in the 24-dimensional features.

In addition to the DFT features, we also experimented with other types of GPS features such as using simple speed, acceleration and turn angle features for individual points. Also, short (2–20 point) windows of relative speed and turn angle were considered. However, the DFT features provided the best performance and is therefore the only one included in this study.

3.2 Gaussian Mixture Model Classifier

We approach user classification of GPS trajectory data as a biometric authentication task following Bayes' decision theory [2]. Given a route $\mathcal{R}$ represented by a sequence of feature vectors, $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$, and an identity claim of user $u \in \{1, 2, \ldots, U\}$, we evaluate log-likelihood ratio:

$$\text{score}(\mathcal{X}, u) = \log \frac{p(\mathcal{X}|\text{user} = u)}{p(\mathcal{X}|\text{user} \neq u)}. \tag{1}$$

Here the numerator models the target user hypothesis while the denominator models its complement. We assume independent observations $\{\boldsymbol{x}_t\}$ and model both the target and anti-hypotheses using Gaussian mixture models (GMMs),

$$p(\mathcal{X}|\boldsymbol{\lambda}) = \prod_{t=1}^{T} \sum_{m=1}^{M} P_m \mathcal{N}(\boldsymbol{x}_t|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \tag{2}$$

where M is the number of Gaussians (model order) and $\boldsymbol{\lambda} = \{P_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m : m = 1, \ldots, M\}$ denotes the model parameters: mixing weights (component priors) P_m, mean vectors $\boldsymbol{\mu}_m$ and covariance matrices $\boldsymbol{\Sigma}_m$. In our implementation

the covariance matrices are constrained to be diagonal ones[1]. The number of Gaussians, M, is a control parameter to trade off between precise user modeling and generalization power (this will be explored below).

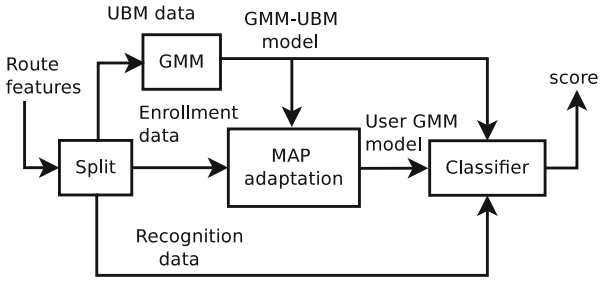

Fig. 3. Modeling and classification system

To train the model, we follow a well-known two-step recipe based on *adapted Gaussian mixture models* or GMM-UBM [12], based on *maximum a posteriori* (MAP) adaptation principle [5] (see Fig. 3). First, a *universal background model* (UBM), used for modeling the anti-hypothesis in (1), is trained by pooling route feature vectors from a large number of off-line users. This is achieved using *expectation-maximization* (EM) algorithm. The UBM is intended for representing common information shared across all the users and is trained once only, using users disjoint from the target users. The UBM is then used in (1) across all the target users for score normalization. To enroll (register) a new user, the UBM parameters are moved towards the enrollment data. We copy the mixing weights and component variance vectors from the UBM and use the adapted mean vectors to represent users. We point the interested reader to [12] for further details.

4 Experimental Setup

4.1 Data Sets

We created four different evaluation protocols based on two publicly available source GPS data sets: **Geolife**[2] and **Openstreetmap**[3]. As the original datasets

[1] This assumption, in general, is non-restrictive. As pointed in [11], since the individual Gaussians act together to model the input feature density, full covariance matrices are not needed even for features that are statistically dependent. For a given amount of training data, one can fit either a GMM with full covariance matrices with less Gaussians, or equivalently a larger GMM with diagonal covariances.

[2] http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/.

[3] http://planet.openstreetmap.org/gps/gpx-planet-2013-04-09.tar.xz.

were originally not designed for biometric tasks, we designed specific evaluation protocols similar to those used in speaker verification and other biometrics. The details of the resulting datasets are presented in Table 1.

Table 1. Data sets.

	OSM30m	OSM60m	Geolife30m	Geolife60m
Target users	156	51	34	20
Users for UBM	1178	793	35	41
GPS sampling interval	1 s	1 s	2 s	2 s
Test segment size	30 min	60 min	30 min	60 min
Test data/user	2 h	4 h	2 h	4 h
Training data/user	2 h	4 h	2 h	4 h
Genuine trials	624	204	136	80
Impostor trials	96720	10200	4488	1520

The Geolife data set was collected by Microsoft Research Asia during a course of three years (from 04/2007 to 08/2012). It contains data from 182 users (which our filtering reduced to 34) logged in varying sample rates, covering a broad range of outdoor activities including life routines, sports activities, shopping, sightseeing, dining, hiking and bicycling.

Openstreetmap (OSM) data is a collection of public traces[4]. A 23 GB compressed dump of these traces was published in 2013[5]. It contains GPS traces uploaded to `openstreetmap.org` by 41413 different users logged in varying sample rates. We select those routes which (1) privacy option was set to "identifiable"[6], and (2) route sampling interval was 1 s. Due to computational limitations, only a subset of these routes were processed.

To reduce the risk of having two almost identical routes (such as from work to home), and thus detecting a route instead of user, each user's routes were filtered by removing any overlapping points (< 30 m). To remove most car routes, we applied a simple heuristic speed filtering by taking the top 4-quantile speed of a segment and discarding the route if the speed exceeded 35 km/h. The route data was then processed to contain only uniform interval (1 s or 2 s) routes. For the 2 s routes, it contains also 1 s routes with every second point removed. Only the OSM data set contained enough data to create a 1 s interval data set.

From the filtered routes, we created a subset where each user has exactly the same amount of training and test data, for example 4 h of testing data split into four 60 min test items for OSM60m. If the desired amount of data (time) was not reached, we excluded the user from the data set. The routes of users

[4] https://www.openstreetmap.org/traces.
[5] http://planet.openstreetmap.org/gps/.
[6] http://wiki.openstreetmap.org/wiki/Visibility_of_GPS_traces.

which did not have sufficient training or testing data were retained for training the universal background model.

4.2 Evaluation

Classifier scores were computed for all the possible (usermodel, route) pairs; whenever the user identity of the test route matches the target (model) user, this constitutes a *same user* (genuine) trial, otherwise a *different user* (impostor) trial. We measure the performance using a standard performance measure of biometric systems, *equal error rate* (EER), which is the misclassification rate at the detection threshold where false acceptance and false rejection rates are equal. In practice, we use an implementation in BOSARIS toolkit[7] to compute EERs. In addition to EER, we also report *relative rank* (RRANK) measure, defined as the average rank of the correct user model score and normalized to range from 0 (best) to 1 (worst) by dividing with the number of users.

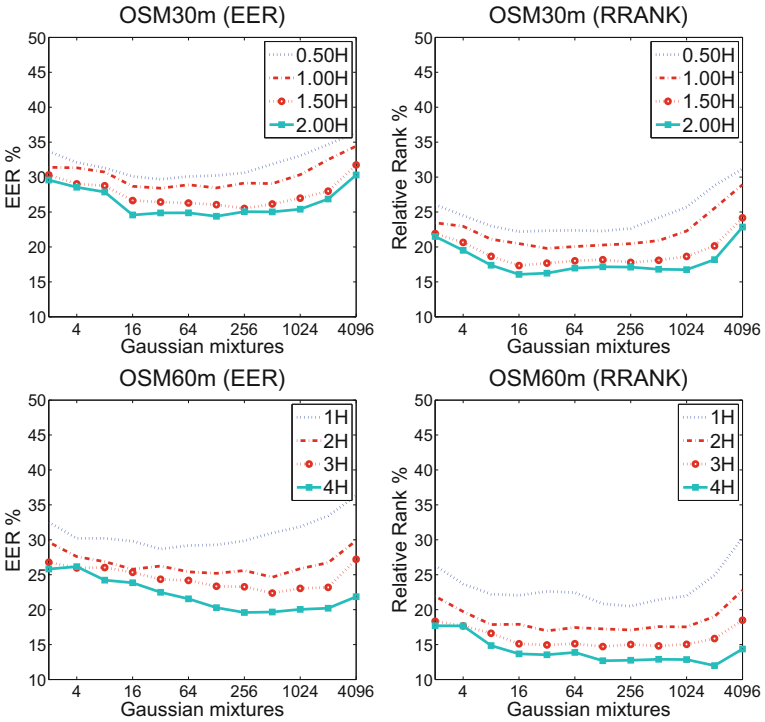

Fig. 4. Results for OSM30m and OSM60m data sets. The amount of training data is varied from 1 to 4 h.

[7] https://sites.google.com/site/bosaristoolkit/.

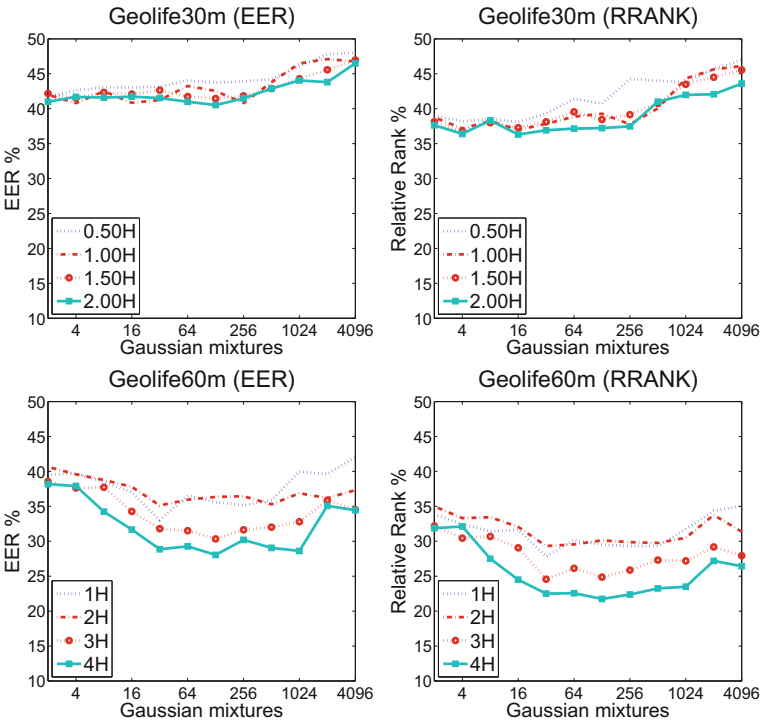

Fig. 5. Results for Geolife30m and Geolife60m data sets. The amount of training data is varied from 0.5 to 2 h.

5 Results

The results are presented in Figs. 4 and 5. The amount of training data is varied from 1 h to 4 h for OSM60m and Geolife60m data sets and from 30 min to 2 h for OSM30m and Geolife30m data sets. The number of Gaussians is varied from 2 to 4096. Best accuracy of 19.6 % EER and 12.8 % RRANK was achieved in OSM60m data set with 256 Gaussian mixtures and 4 h training data. For OSM30m data set best accuracy of 24.4 % EER and 17.2 % RRANK was achieved with 128 Gaussian mixtures and 2 h training data.

Comparing the two datasets, the recognition accuracy is lower for the Geolife data. For Geolife60m, the best accuracy (28.0 % EER and 21.8 % RRANK) was achieved with 128 Gaussians and 4 h of enrollment data. For Geolife30m, the best accuracy (40.5 % EER and 37.2 % RRANK) was achieved with 128 Gaussians and 2 h training data. Three possible reasons for the lower accuracy for Geolife include (1) lower GPS sampling, leading to less discriminative spectral features, and (2) much smaller number of users to train UBM.

Concerning the amount of training data and the number of model parameters, we observe three expected results. Firstly, larger amount of training data generally leads to higher accuracy. Secondly, the optimal number of Gaussians lies

in between the tested parameter range. This is expected from the bias-variance trade-off in statistical modeling: too many Gaussians leads to overfitting while too few do not discriminate the users well. Thirdly, for larger amounts of training data, the optimal model size is obtained with a larger number of Gaussian components.

Compared with other user movement based biometrics, the GPS features did not achieve as good recognition accuracy as accelerometer. For example, the gait-based recognition in [9] reached 7 % EER, compared to our 19.6 % using GPS signal. Although it is only indirect comparison, it is reasonable evidence that even if GPS signal can be used to recognize user, accelerometer probably provides more reliable source—if available. GPS technology is also likely to develop further to make it more reliable. Probably also more accurate in user identification. One potential future idea would be to study the joint use of GPS and accelerometer data.

6 Conclusion

Our experiments indicate that local variations in GPS data possess user specific characteristic that can potentially be used to recognize a person and should therefore be handled with similar care as any other private information such as voice or fingerprints.

Our method achieved an accuracy of 19.6 % EER and 12.8 % RRANK in the best case. While these error rates are clearly too high to be useful in user authentication applications requiring high level of security and trustworthiness, they do indicate that local GPS trajectory movements contain person-identifying information. This information might be useful for applications such as recommendation systems or to detect sudden changes in user's behavior.

We assume that the accuracy is less than optimal partly due to inherent inaccuracy of GPS data and low sampling frequency on available data sets. However, future development in movement tracking technology is likely to increase accuracy of routes, and consequently improve recognition accuracy of our method. Also, even larger amount and diversity of training and testing data would likely improve the accuracy further.

Due to limitations of available data sets, we were unable to rule out the impact of certain factors which may be user specific while not being characteristic of users. These include properties of GPS tracker, and the area where user lives. Therefore it is still an open question how much the recognition accuracy relates to what user is and how much to user's surroundings.

References

1. Biagioni, J., Krumm, J.: Days of our lives: assessing day similarity from location traces. In: Carberry, S., Weibelzahl, S., Micarelli, A., Semeraro, G. (eds.) UMAP 2013. LNCS, vol. 7899, pp. 89–101. Springer, Heidelberg (2013). doi:10.1007/978-3-642-38844-6_8

2. Duda, R., Hart, P., Stork, D.: Pattern Classification. WIS, New York (2000)
3. Fletcher, D.: Please rob me: The risks of online oversharing. Time Mag. Online (2010)
4. Gambs, S., Killijian, M.O., del Prado Cortez, M.N.: Show me how you move and I will tell you who you are. In: Proceedings of 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS, SPRINGL 2010, pp. 34–41. ACM, New York (2010)
5. Gauvain, J.L., Lee, C.H.: Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. Speech Audio Process. **2**(2), 291–298 (1994)
6. Krumm, J.: A survey of computational location privacy. Pers. Ubiquit. Comput. **13**(6), 391–399 (2009)
7. Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., Ma, W.Y.: Mining user similarity based on location history. In: Proceedings of 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS 2008, pp. 34:1–34:10. ACM, New York (2008)
8. Liu, H., Schneider, M.: Similarity measurement of moving object trajectories. In: Proceedings of 3rd ACM SIGSPATIAL International Workshop on GeoStreaming, IWGS 2012, pp. 19–22. ACM, New York (2012)
9. Mäntyjärvi, J., Lindholm, M., Vildjiounaite, E., Mäkelä, S.M., Ailisto, H.: Identifying users of portable devices from gait pattern with accelerometers. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. ii-973. IEEE (2005)
10. de Montjoye, Y.A., Hidalgo, C.A., Verleysen, M., Blondel, V.D.: Unique in the crowd: the privacy bounds of human mobility. Scientific reports 3 (2013)
11. Reynolds, D.: Gaussian Mixture Models. In: Li, S.Z., Jain, A. (eds.) Encyclopedia of Biometric Recognition. Springer, New York (2008)
12. Reynolds, D., Quatieri, T., Dunn, R.: Speaker verification using adapted gaussian mixture models. Digit. Sig. Process. **10**(1), 19–41 (2000)
13. Rossi, L., Walker, J., Musolesi, M.: Spatio-Temporal Techniques for User Identification by Means of GPS Mobility Data. CoRR abs/1501.06814 (2015)
14. Song, C., Qu, Z., Blumm, N., Barabási, A.L.: Limits of predictability in human mobility. Science **327**(5968), 1018–1021 (2010)
15. Waga, K., Tabarcea, A., Chen, M., Fränti, P.: Detecting movement type by route segmentation and classification. In: 2012 8th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), pp. 508–513. IEEE (2012)
16. Wang, H., Liu, K.: User oriented trajectory similarity search. In: Proceedings of ACM SIGKDD International Workshop on Urban Computing, UrbComp 2012, pp. 103–110. ACM, New York (2012)
17. Ying, J.J.C., Lu, E.H.C., Lee, W.C., Weng, T.C., Tseng, V.S.: Mining user similarity from semantic trajectories. In: Proceedings of 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, LBSN 2010, pp. 19–26. ACM, New York (2010)