Processing Mobility Traces for Activity Recognition in Smart Cities

Arsalan Shah, Petr Belyaev, Borja Ramis Ferrer, Wael M. Mohammed, Jose L. Martinez Lastra
Tampere University of Technology, Tampere, Finland
{syed.a.shah, petr.belyaev, borja.ramisferrer, wael.mohammed, jose.lastra}@tut.fi

Abstract— **Human mobility modelling has emerged as an important research area over the past years. The opportunities that mobility modelling offers are widespread. From smart transportation services to reliable recommendations systems, all require generation of mobility models. Since mobility of humans is generally motivated by the activities they perform, activity recognition emerges as a vital initial step towards building better and accurate mobility models. The activity recognition can be carried out by analyzing relevant data from GPS devices, accelerometers and many other sensing sources. The most common approach is to combine data from different sources, analyze that data and recognize the type of activity being performed. However, this requires access to many specialized devices and customized infrastructures. As an alternate, this paper introduces a novel approach to recognize activities from the GPS traces only. This approach utilizes Adaptive-Neuro-Fuzzy Inference System (ANFIS) which combines the power of neural networks and fuzzy logic to recognize activities. The approach is tested on three different datasets and shows promising results. In addition to this a multi-cloud architecture is proposed, for the deployment of such a system.**

Keywords—mobility traces; activity recognition; smart cities; data mining; adaptive-neuro-fuzzy inference system

I. Introduction

The enhancements of telecommunication infrastructures over the last decades permits the collection of large amounts of data from inhabitants of cities. Generally, such data does not provide useful information because it is received in unprocessed and raw formats. However, it is possible to discover many patterns that lead to implicit and valuable information from the collected data. The steady decline of computing cost [1] has resulted in improvement of methods and techniques, that can be employed for exploration of this implicit information.

Mobility traces are pieces of data that can be used to reconstruct the path of motion of individuals. This information can be collected in several forms, such as Global Positioning System (GPS) coordinates or Global System for Mobile (GSM) cell identifiers. When combined with timestamps, mobility traces can provide valuable information about an individual This information is not only valuable for that individual, but also for organizations in the commercial and the public sector. In fact, due to its value, data collectors may decide to use the data themselves or sell it to third parties.

One of the potential benefits of using mobility traces of the inhabitants is to enhance transportation services of cities. For example, as the location of people at specific times of the day can be inferred, mobility models can be built and, afterwards, employed for predicting traffic jams before they occur. Similarly, discovered mobility patterns can be a powerful asset for increasing the efficiency of public transportation. In addition, the information retrieved from such mobility traces can be used to identify potential locations for building commercial centers, businesses and restaurants. Conceptually, it is possible to identify groups of people based on their interests and bring better services to the locations wherein they spend more time along the day. This requires efforts on analyzing mobility traces and finding patterns.

However, extracting patterns only from a set of timestamped mobility traces is a difficult task. Therefore, it would be useful to have more information about the traces. For example, if an individual spends their evenings in a coffee shop or in a fitness center, it is hard to incur such information only from the mobility traces. Hence by only considering timestamps and traces we lose valuable information about the place that is being visited. Having semantic information, such as the motivation of visiting a place can greatly enhance the understanding about specific places and individuals.

To some extent, the rise of social media [2], where platforms (e.g. Facebook[1] and Twitter[2]) provide their users the opportunity to check-in to a place or record their thoughts and feelings has lightened focus from the problem of obtaining semantic information from mobility traces. But these platforms provide limited access to the data they collect. Then, it could be claimed that it is relatively easier to collect only mobility traces, rather than collecting mobility traces tagged with semantic information. Nevertheless, the collection of only mobility traces makes difficult to build accurate and useful mobility models.

This research work focuses on extracting semantic information from the heap of mobility traces. The presented approach proposes the combination of fuzzy reasoning with neural networks for obtaining semantic information from timestamped mobility traces only. This information describes the activities that an individual undertakes at specific times and locations during the day. Therefore, this article presents an approach for activity recognition that, indeed, is useful in two ways. Firstly, it helps for offering better services to the people. Fundamentally, this is achieved with recommendations that can be made for trying out similar activities offered by a different

[1] https://www.facebook.com/
[2] https://twitter.com/

service provider. Secondly, the approach permits the creation of semantic maps through the collected data from multiple individuals. In turn, these maps can provide additional information about areas for which any prior information is not available. Then, this may support the development of systems which are independent of other third-party APIs such as Google Maps[3].

The rest of the paper is structured as follows: Section II presents a literature review and current practices on activity recognition, and cloud based systems for smart environments. Then, Section III presents the approach for processing mobility traces for activity recognition in smart cities. Then, Section IV presents a use case for proving the concept. Finally, Section V concludes the paper and describes some further work to be performed.

II. Literature Review and Current Practices

A. Activity Recognition

The dynamics of people's mobility is greatly dependent on performed activities. Although the activities are systematic and organized from the point of view of the subject, to an observer that is not aware of the subjects' nature and routine, these mobility patterns might appear completely random. Therefore, previously modelling human mobility was considered as a stochastic problem [3] rather than a deterministic one. However, as shown in [3], human mobility follows patterns that can have a potential predictability up to the 93%. This value has interesting and useful implications. For example, as discussed in previous section, urban infrastructures can use activity recognition for improving their services. Thus, motivated by such applications researchers have tried to extract patterns in human mobility.

Many methodologies have been put forward to identify mobility patterns, construct models and predict the mobility behavior of people individually or as a group. It is possible to construct these models by considering only the temporal or geospatial aspect as the basis for the movement. But mostly the combination of both space and time is used to generate better models of the mobility. For instance, the research works [4]–[6] use this combination of spatio-temporal information for analyzing mobility patterns. Many other researches, such as [7]–[11] have used the same spatio-temporal aspect for building next place prediction systems. Even though building an accurate next place prediction system is the ultimate goal to be achieved after the results presented in this research. Such a system is out of scope of the current work, where the focus is on activity recognition only. In [4] the authors argue that human mobility is dependent on the intentions of the movement and by incorporating these intentions in our analysis we can build better mobility models. Based on their argument they have categorized these intentions into three sub-categorizes; namely Geographically-triggered Intentions (GI), Temporal-triggered Intentions (TI) and Semantic-triggered Intentions (TI). Since the research works described in [4], [5] are based on spatio-temporal analysis of mobility patterns, they are able to capture GI and TI but fail to grasp SI. The idea behind an SI is that such intentions reflect why someone wants to travel from one location to another. For example, if a person goes to a restaurant after working hours, having semantic information about the two locations can enrich the mobility model by adding extra information to it. It can be seen that in the case of SI it is important that the nature of activity being undertaken at a certain location must be known in order to add the extra contextual information. In such a case, obtaining the semantic information can be regarded as identifying the activity.

Identification of activities is a problem where the method of identification is greatly dependent on the type of data available. Most of the approaches for activity recognition that are relevant to this research work use data collected from sensors attached to or in close proximity of a user. The data can be made available from accelerometers [12], [13] state-change sensors [14] or other wired or wireless locating devices [15]. Activity recognition based on these sensing devices shows promising results, but it poses the challenge of collecting this data. Special sensors need to be installed for collection of this data, which limits the number of users of the devices. Recording accelerometer data from mobile phones is possible but that is also not openly accessible for analysis. Smart phones have made it significantly easier to collect location data from its users. Location information from any GPS containing smart device is relatively easier to obtain and more accessible to researchers. Some researchers [16] have also combined GPS data with accelerometer data to identify activities, showing good results but again making the system rely on the more private accelerometer data.

Some research work towards identification of activity patterns from spatio-temporal data only, can be found in the literature. For example [17], [18] use GPS trajectories for activity recognition, while [19] uses a combination of GSM cell identifiers and GPS coordinates to find important locations of users.

Furthermore in the work of [20] the emphasis is on extracting the activity patterns when the activities are pre-identified. On the other hand, [11], [18] use an existing database of landmark tags. This requirement of prior information puts these methodologies at a major disadvantage. Currently, it can be claimed that [18], [19] are the only works done for identifying interesting locations or activities from location data only. Even though [19] gives a method for identifying interesting locations and not the actual activities, the type of activities can be inferred from the classification of the identified locations. The authors demonstrate the use of various clustering techniques to extract and identify similar locations or activities. On the other hand, [18] use Markov networks along with other techniques for activity recognition. Moreover, the research work presented in [17] closely relates to this research work, wherein machine learning techniques are used for activity identification. However, this research work presents a novel approach where an Adaptive Neuro-Fuzzy Inference System (ANFIS) is used to identify activities of users solely based on GPS trajectories. Such systems combine the power of Neural Networks and Fuzzy Logic for providing a relatively simple and reliable way of activity recognition.

[3] https://developers.google.com/maps/

B. Cloud based systems and security (MUSA)

Data acquisition is frequently done using some kind of web service. During the last decade, web service security became one of the primary concerns of information security around the world. Multiple attack vectors can be used to not only crash the service, but to compromise and exploit user private data delegated to it. Some of those vectors can be found in the most recent Open Web Application Security Project's (OWASP) "Top 10" publication [21], which provides both explanation of attacks and methods how to secure the application from those.

Moving such service into the cloud and, especially, distributing it between multiple clouds to improve availability and end-user latency, makes its security even more vulnerable. Monitoring and enforcing security in the cloud might be difficult not only in terms of mechanisms [22], but in terms of the agreement with the cloud service provider (CSP). This part is especially hard, as the CSP might not want to provide some of the necessary security controls to the client, and there is no CSP comparison list at the moment [23].

The only manner to ensure the application security while dealing with the cloud is to consider the deployment model and its issues from the first stages of implementation. The ongoing MultiCloud Secure Applications (MUSA)[4] project is a part of EU H2020 programme that provides a set of tools, as well as an application design methodology [24] to ensure effective threat mitigation. The core application of the MUSA project follows the Agile methodology to help developing, deploying and monitoring multi-cloud applications. The presented research work manipulates and processes sensitive data that must be collected from several data sources on smart cities. Hence, employment of solutions provided by MUSA project will ensure both high-performance and security.

III. Processing Mobility Traces for Activity Recognition

The goal of this paper is to present an approach for recognizing the activities of individuals at certain locations of specific environments. This is possible by processing mobility traces that may provide useful information about activities of individuals. This section includes a set of architectural components, their interactions and the structure of the data to be processed.

A. Architectural components

The main objective of this publication includes presenting an approach for identifying the daily activities for a user using ANFIS model. Besides, a visionary deployment is presented using the architecture of the MUSA project. Such project aims to provide a decentralized security for web-based applications. As depicted in Fig. 1, the MUSA platform appears as a set of tools in the cloud. In the context of this research work, the server side contains three main modules. Firstly, the Identity Manager (IDM), which manages the access, control for the users. This authentication employs user credentials like username and password. Besides, a token might be used for session expiration control. Secondly, the Tampere Smart Mobility (TSM) Engine deploys the ANFIS model for training and predicting purposes. As well, it contains the communication interface, which in this case is an HTTP[5] service interface. Finally, a database repository is used for storing the ANFIS model.

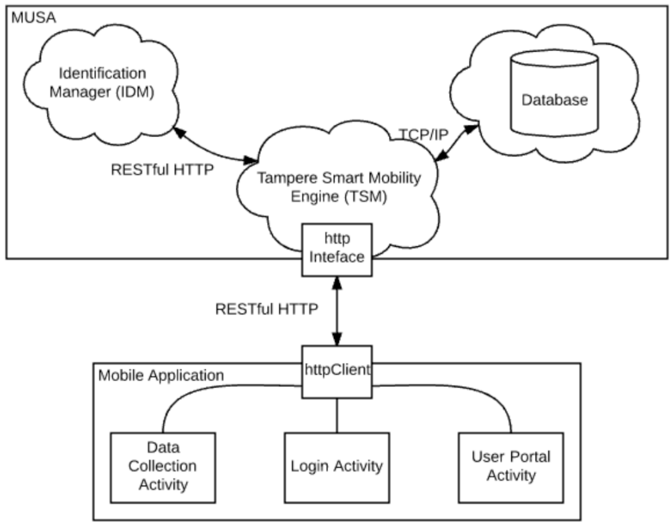


Fig. 1. Architectural view for the activity recognition approach in smart cities

On the other hand, the mobile application contains four main components as follows: i) HTTP Client, ii) Data Collection Service (DCS), iii) Login Activity and vi) User Portal Activity (UPA). The DCS collets the activities of the user with respect to the geolocation and time. This set of data is required for training the ANFIS model as it is illustrated afterwards. Then, the Login Activity, which allows the user to provide the necessary credentials for the authentications process. Meanwhile, the UPA allows the user to interact with the backend application, which is deployed in the MUSA platform. Finally, the HTTP client serves the mobile application with the necessary communication that is needed from the application.

B. Interaction of components

The previous subsection presented an architectural view of deploying the ANFIS model for tracking the users' daily activity by exploiting the MUSA platform features. This subsection tends to highlight the interaction between the presented components. As seen in Fig. 2, initially, the sequence starts once the user login using the mobile application using the Login Activity in the mobile. Accordingly, the Login Activity creates the authentication information and then, send it to the TSM engine. Consequently, the TSM engine verifies the user by sending the authentication information to the IDM. Finally, the IDM responds with acknowledgement to the TSM, which passes it to the Login Activity. The acknowledgement holds the results of the authentications. Besides, it might contain a token for controlling the session.

[4] http://www.musa-project.eu/

[5] https://www.w3.org/Protocols/

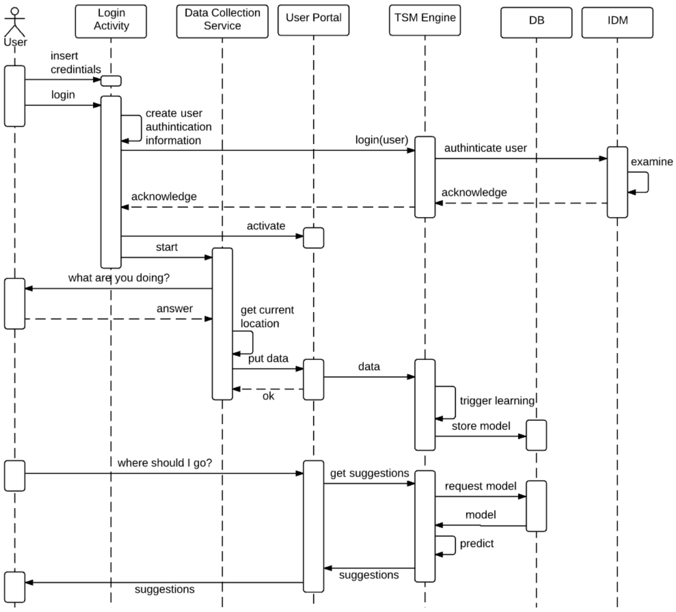


Fig. 2. Interactions sequence diagram

Once the authentication is accepted, the Login Activity starts the DCS and UPA. The DCS collects the activity of the user with respect to the geolocation and time. This routine could be occurred on time bases e.g., in hourly bases, or it can be according to the changes to the location of the mobile itself. Another approach could be using the calendar of the user. This data represents the learning material that the ANFIS model requires for the prediction purposes. According to the sequence diagram in Fig. 2, the collection of the data starts by asking the user about what he/she is doing. After collecting the daily data, the DCS sends this data as a PUT RESTful service. This service is handled by the TSM engine. Concurrently, the TSM starts the learning process of the ANFIS model. After that, the TSM stores the model back to the databases. This process could be repeated each day since the mobile application sends daily data to the TSM.

In the meantime, the user could get suggestions about the activity that he/she might need to do. This requires the user to use the UPA to request suggestions. Then, this request is transferred to the TSM, which uses the trained ANFIS model to predict the required activity. Finally, the TSM responds with the suggested (predicted) activity that the user might be interested in. It is important to mention that this example interaction is a visionary implementation and can be extended to a standalone component, e.g. an intelligent recommendation system. This component can then be integrated into other application software to improve the user experience.

C. Model

In order to identify activities from the dataset, a fuzzy based approach is used. One of the advantages of using fuzzy inference system is the ability to have linguistic variables. The linguistic IF-THEN rule based system mimics the human decision making and should be able to identify activities if enough information is presented to such a system. The model is trained with two inputs, starting time of the activity and the time spent performing that activity. The output is a set of four variables, representing the four activities that are considered for this study.

The motivation for using only two inputs comes from imagining what would be the most relevant information required for a human being to recognize activities. The most relevant information for such a situation will be to know at what time of the day an activity was started and how long did that activity last for. Based on this knowledge, basic information can be collected on what activity is being performed. There can be more inputs to the model such as day of the week or time of the year but this information will affect the model in terms of where the data was recorded. For example, a week may start from a different day in some countries than the others. Thus, in the effort to make the model generic enough these inputs are not considered.

The four activities that are selected as the output from the system are the activities that are performed on a daily basis. These include activities done at home, the most common being sleep or rest, such activities are labelled as “home”. The activities performed at work or at school are labelled as “work”. The activities that are done on a routine basis but which do not fall under work or home are labelled as “chores/leisure”. These activities include shopping and other activities performed outside the house such as going to the gym or a night out with friends. The fourth activity label is “eating” and this will cover outdoor dining at restaurants or cafes etc. These activity labels cover the most common tasks that an individual may perform during the day and, repeatedly, during the week. More activities can be added or removed from this set as per requirement. But it is not possible to make an activity list that encompasses all the activities a human being may undertake. Work of [25] can be used as a guideline to decide which activities to consider.

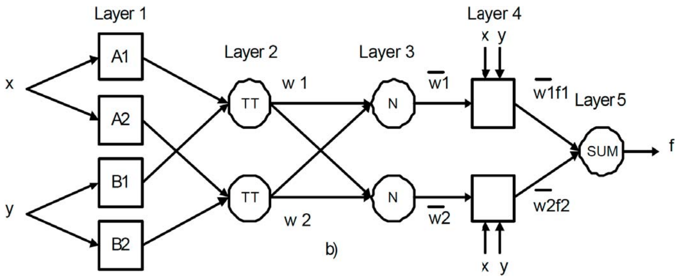


Fig. 3. A typical ANFIS structure for two inputs, two rules and one output [26]

Fig. 3 shows a typical ANFIS network where the output f is given by equation 1 [26].

$$f = \frac{w_1}{w_1 + w_2} f_1 + \frac{w_2}{w_1 + w_2} f_2 \qquad (1)$$

IV. Proving the Concept

The main objective of the publication is to identify and recognize activity patterns from a data set of raw timestamped GPS traces. As this work is a proof of concept so the activity recognition algorithm has been developed in Matlab[6]. After analyzing the results and getting satisfactory outputs we will

[6] https://www.mathworks.com/

move the algorithm will be implemented in a software capable of easy deployment.

A. Use case

To test the performance of the activity recognition system, this research work employed real data that contains mobility traces of a diverse set of individuals. More precisely, two open datasets are selected. The first dataset was collected in Geolife project of Microsoft Research Asia [5], [27], [28]. This dataset is collected from 182 users over a 5-year period. The data is recorded using GPS loggers and GPS phones. The amount of GPS traces per user and data sampling rates vary. The second dataset is extracted from Mobile Data Challenge (MDC). This dataset contains timestamped GPS data from 141 users collected over a period of one and a half year from Switzerland. The GPS data in MDC dataset is collected from mobile phones. Like the Geolife dataset the sampling rate of this dataset is also not constant. Both the datasets are anonymized before distribution so the users are identified by a user ID only.

There is other information also present in these datasets e.g. speed, altitude, mode of transportation etc. Since this research work experiment is only interested in the Latitude, Longitudes and timestamps, a small Matlab subroutine is written to clean the dataset of all extra information and convert timestamps to UNIX if required.

In addition to the aforementioned datasets, location history of some researchers of this research group is also used in the study. This data was generated by first allowing Google to record the GPS data and later downloading the recorded data. Google[7] provides the data in the form of a JSON [29] or KML[8] file. The data recorded can be rich with information irrelevant for this study. Again, a small Matlab subroutine is used to extract relevant data from the JSON file.

B. Implementation and testing

The algorithm for extracting activities from the raw GPS trajectories is a three-step process. In the first step, all the GPS traces and timestamps are processed to extract the stay points. In fact, this research work employs the approach described in [30] in order to extract the stay points. The stay points can be defined as a set of GPS traces which lie within a radius of X meters of each other and the time spent from the first till the last trace is at least Y minutes. For the experiment, 30 meters has been chosen as the value of X and 20 minutes as the value of Y. As a result of this step, stay points are obtained. Each stay point is represented by a center point p and an effective activity radius r. The center point is the mean of all the GPS trajectories forming that specific stay point and the radius is the distance from p to the farthest trace within the stay point. In this study one stay point corresponds to one activity.

In the second step consists of combining similar activities of a single user are combined. One requirement for the two activities to be considered similar is that the distance between the centers of the activities should be less than a specific distance. In performed calculations, this number has been fixed at 10 meters. The other condition is that the starting time t_s and ending time t_e of the two activities should not be more than one hour apart.

[7] https://takeout.google.com/settings/takeout
[8] https://developers.google.com/kml/

Algorithm 1 Find stay points and group similar activities

Input: GPS trajectories
Output: Activities

```
For all users
    Find all the activities (stay points) from the traces, where each stay
    point is defined by a center p and radius r
    For all activities
        If p_n - p_{n+1} is less than 10m and abs (t_sn – t_sn+1) is
        less than 1hr and abs (t_en – t_en+1) is less than 1hr
            merge activities (n,n+1)
        end If
    end For
end For
merge similar activities
```

Fig. 4. Algorithm for mining activities from mobility traces

After the activities from one user are identified, the same algorithm is repeated for all the users. In the final step the activities from different users that are similar are grouped together. Fig. 4, gives the algorithm that is used for extracting the activities. While merging similar activities of one or all users together, it is observed that the radius of activities can grow as two activities can have different effective activity radius r. This creates problems because the presented approach aims to recognize activities in urban areas where many services can be provided in a short space. Hence, too large r values result in inaccuracies. Then, in order to shrink the effective radius of each activity the Chebyshev's inequality of equation 2 [31] is used to calculate a confidence interval for each activity. By setting the probability value of 50%, and calculating variance along the latitude and longitude it is possible to shrink the effective radius and thus point to activity locations more accurately. Fig. 5, shows one example of the use of Chebyshev's inequality to reduce the radius r.

$$P(|x| \geq \varepsilon) \leq \frac{E(x^2)}{\varepsilon^2} \qquad (2)$$

After the activities are identified the information about the start time of each activity and the time it takes to be completed are obtained. Then a trained Adaptive Neuro-Fuzzy Inference model is used to output the predicted activity.

Collecting real data sets with activity labels can be a challenging task and we do not have any labelled dataset. As mentioned earlier this experiment also uses data collected by Google for some of the researchers involved in this study. This data is processed to extract activities using the algorithm of Fig. 4. Since this data belongs to the researchers it is possible to attach activity labels to it. This labelled data is then used to train the ANFIS model. This technique is not very useful since the test datasets are very big and the trained model is only trained for a few inputs and is not able to identify the activities correctly.

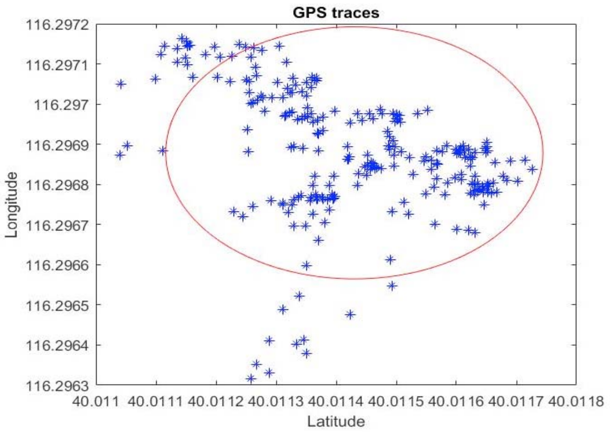


Fig. 5. The 50% confidence ellipse plotted for the traces of one stay point. The new radius is found by calculating the distance between point of intercept of the ellipse on the principal axis and the center

Thus, another approach was used which involves Matlab's fuzzy toolbox. With the help of a Fuzzy controller the task of labelling the activities has been automated. A Mamdani type fuzzy controller has been created and a set of rules are used to identify activities. A total of 13 input member ship functions covering most common scenarios and time of the day are created. These membership functions are used to define 42 different AND/OR rules. Based on these rules activity labels are given as the output variable. The rules are defined using human intuition, e.g. if an activity starts at 10:00 am and lasts for 6 hours it is highly likely that this activity is related to work. The rules and membership functions are defined in such a way that each activity label gets the output value based on the likelihood of that activity happening. Fig. 6 shows membership functions for the first input variable *time of the day* and Table 1 contains sample of the rules that are used. In the end, this whole process provided activity labels for the dataset.

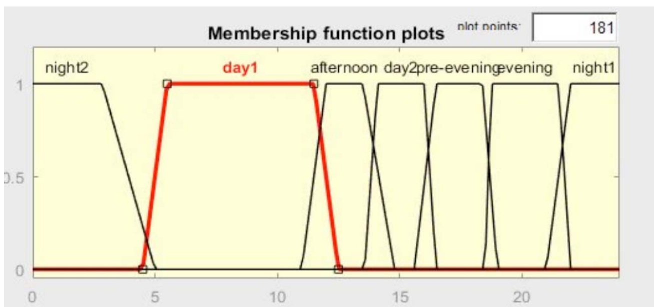


Fig. 6. Membership functions for the first input variable *time of the dayp*

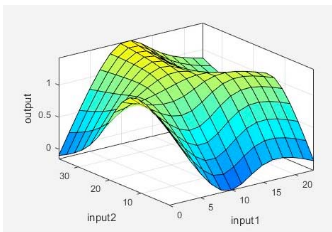
Table 1. Example of rules for the fuzzy controller

No	Rule
1	If (TimeOfDay is day1) and (HoursSpent is v.less) then (home is no)(work is no)(leisure is no)(eating is yes) (1)
2	If (TimeOfDay is day1) and (HoursSpent is many) then (home is no)(work is yes)(leisure is no)(eating is no) (1)

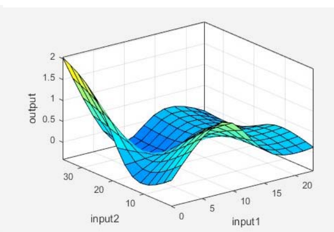
The accuracy of the output labels is verified visually with the help of Google Maps API. Once a satisfactory output is obtained, this labeled dataset is then used as training data for the ANFIS model. After training the ANFIS model with this new training data the model is tested with the other dataset. The whole process is done such that if the ANFIS model is trained with Geolife dataset then it is tested with MDC and google data, and if the training is done with MDC dataset the testing is carried out on Geolife and google dataset.

Since the trained model is used to output four different activities so four different ANFIS models are required, one for each activity type. The advantage of this setup is that the probability of an activity belonging to each class can be known. Then, in this way an activity is not classified in a binary manner, but assigned some likelihood value. In the end, all the individual ANFIS model outputs are combined. Since this classification method assigns probability values to all activities, a weighted average is taken before tagging or labelling the activity stay point. This averaging takes into account the number of visits which were labeled as activity 1 and number of visits which were labeled activity 2 and so on. The visits are multiplied with the output probability and then normalized. By averaging in this manner, the activities that are performed frequently get more weightage in the assignment of labels.

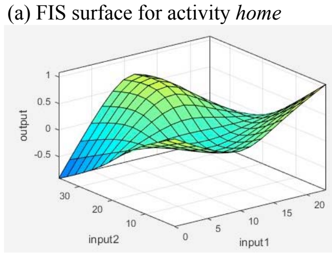
Once this information is processed, labels are added to activities that have a weighted average value greater than 0.5. In this way one stay point might have more than one activity label. The Fig. 7 shows the surface plots for the four ANFIS models that are used for activity identification. As can be seen from the Fig. 7 (a), (b), (c) and (d) all the fuzzy inference system (FIS) structures are non-linear but exhibit smooth surfaces with no abrupt gradients. This shows that the trained models adapted well to the input data but it also shows the absence of outliers which are not present due to the auto generated labels. One important observation is the linear behavior of surfaces closer to the edges, while the presence of bumps in the middle. This behavior is indicative of the challenge in differentiating activities when they are performed in the middle of the day and last less than 8 hours.



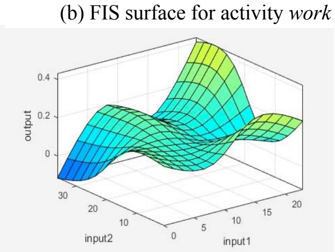
(a) FIS surface for activity *home*



(b) FIS surface for activity *work*



(c) FIS surface for activity *leisure*



(d) FIS surface for activity *eating*

Fig. 7. Surface plots for different Fuzzy Inference Systems

C. Results

No publicly available dataset was found which contains labeled activities where the working of the algorithm could be verified. Then, the datasets that are used for testing do not contain any relevant contextual information. To overcome this problem the activities and their respective labels are plotted on the Google Maps to visually verify the results. For visualizing the results an HTML[9] page is created where the results are plotted on top of Google Maps with the help of a Javascript code. Here Fig. 8 shows results for the MDC dataset. Since the MDC dataset is collected from a large number of people it is only possible to check the results visually to some extent. It can be seen from the Info window in Fig. 8, that the activities with labels leisure or eating include areas that are restaurants, grocery stores or some recreational spots like parks etc. The places that are labeled as home are mostly residential areas. Such observations add some credibility to the model. Since the Google data is collected from the researchers in the research group the results for that data can be verified with more certainty. It can be seen from Fig. 9 that the trained ANFIS model is successful in identifying home, work, leisure and eating. This is verified by showing the results to the researchers and asking them if the output labels are true.

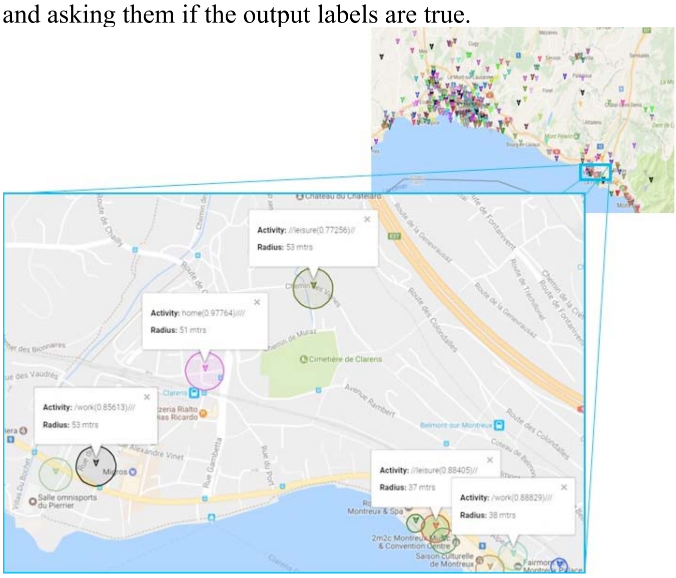


Fig. 8. Visualization of result for MDC dataset

The most critical part of identifying activities is the data mining step. The results are greatly affected by the extraction of stay points from the mobility traces and how the similar activities are merged. The first parameters that affects the extraction of stay points is the size of the stay point where an activity takes place, this is defined by *r*. This parameter can vary based on the nature of the activity e.g. a walk in the morning will encompass more area than shopping at a grocery store. Another important thing to consider is that the mobility traces are collected from GPS devices so the collected data is not precise. The precision of the data can vary based on the type of instrument used for collection of data. Hence the cutoff value for distance needs to take both the problems into consideration. Similarly, the time spent performing an activity also varies based on the nature of the activity. Therefore, selecting the parameter *t*, which defines the minimum time spent at a stay point, is also tricky. The values of *r* and *t* also affect the processing time of the algorithm as too small values will make the algorithm very slow and extract false stay points. On the other hand, too large values will lead to ignoring important stay points.

[9] https://www.w3.org/community/webed/wiki/HTML/Specifications

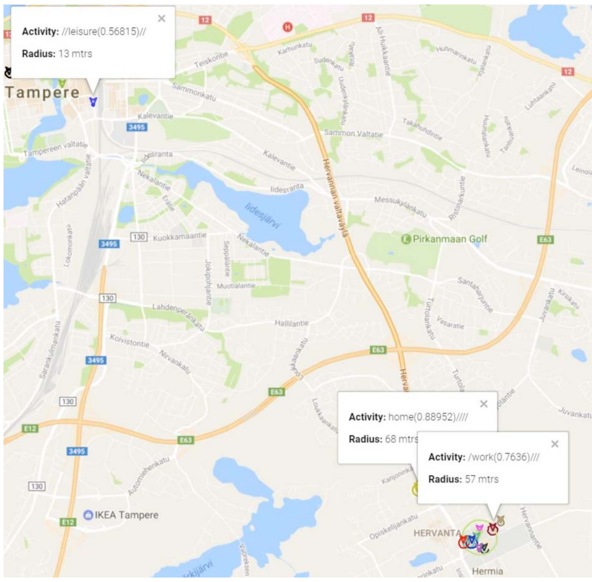


Fig. 9. Visualization of result for Google data of one of the researchers

Before deciding on the parameter values all the aforementioned points are taken into consideration. Finally, stay points are extracted by setting *r* as 30 m and *t* as 20 min. This was decided due to the fact that these numbers provided good results while extracting stay points for the Google data. Another important step is the merging of similar activities. The merging is based on the distance between the centers of the stay points *d*, and the closeness of the stay points to each other with respect to time *t*. For extracting the most number of stay points while avoiding repeated extraction of similar activities the value of *d* and *t* are set as 10 m and 60 min respectively. These numbers are also obtained after conducting many trials on the Google data.

One of the biggest factors that affects the results is the sampling time of collected data. The results are fairly accurate if there are no dark patches in the data, meaning that the data captures all the movements of an individual. The MDC and Geolife datasets present problems in this regard as there are times when no data is available. This leads to poor extraction of stay points. Google data, on the other hand, captures most of the movement of an individual. Hence the activity recognition is more accurate for Google data.

V. Conclusion and Further Work

The use of fuzzy logic for recognition of activities is interesting, as the approach is similar to human decision making. The output of the model is not *true* or *false* and can be thought of as a degree of *truth*. This is useful in situations where one stay point is visited by more than one people for different purposes. In such cases a stay point can be assigned more than one activity labels with a percentage of confidence. Moreover, and advantage of a fuzzy based activity recognition system can be seen where instead of collaborated labelling of activities the system is used for activity recognition for one individual. Since the model presented is simple enough, multiple similar model can be trained for every individual and used for activity recognition. One disadvantage of fuzzy systems is that the ANFIS model might not be able to adapt to big datasets and introduction of outliers will pose a definite challenge. However, since the activity recognition is the initial step of this research and the main motivation is to extract semantic information from mobility traces for improving the prediction of human mobility. This additional information, even if not 100% accurate, will still help in building more accurate mobility prediction systems.

With the help of work presented in this paper the semantic information about the GPS traces can be extracted. This information will be utilized in the future work, where more accurate mobility models will be generated using this extracted information. These mobility models will then be used for making prediction on mobility of people. At the moment, accuracy of activity prediction is determined by visual inspection. Obtaining accurately labelled datasets for performed activities is a challenging task. Nevertheless, the application will have further an interface to record the actual labels from the users which will then help in generating Receiver Operating Characteristic (ROC) curves and give better insight into the predictor.

ACKNOWLEDGMENT

The project leading to this paper has received funding from the European Commission's Horizon 2020 research and innovation programme under grant agreement n° 644429 correspondent to the project shortly entitled MUSA, Multi-cloud Secure Applications. In addition, the research in this paper used the MDC dataset made available by Idiap Research Institute and owned by Nokia.

REFERENCES

[1] W. D. Nordhaus, "The Progress of Computing," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 285168, Sep. 2001.

[2] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Bus. Horiz.*, vol. 53, no. 1, pp. 59–68, Jan. 2010.

[3] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of Predictability in Human Mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, Feb. 2010.

[4] W. Huang, S. Li, X. Liu, and Y. Ban, "Predicting human mobility with activity changes," *Int. J. Geogr. Inf. Sci.*, vol. 29, no. 9, pp. 1569–1587, Sep. 2015.

[5] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from GPS trajectories," in *Proceedings of the 18th international conference on World wide web*, 2009, pp. 791–800.

[6] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, Jun. 2008.

[7] A. Asahara, K. Maruyama, A. Sato, and K. Seto, "Pedestrian-movement Prediction Based on Mixed Markov-chain Model," in *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, New York, NY, USA, 2011, pp. 25–33.

[8] W. Mathew, R. Raposo, and B. Martins, "Predicting Future Locations with Hidden Markov Models," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, New York, NY, USA, 2012, pp. 911–918.

[9] S. Gambs, M.-O. Killijian, and M. N. del Prado Cortez, "Next Place Prediction Using Mobility Markov Chains," in *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*, New York, NY, USA, 2012, p. 3:1–3:6.

[10] D. Ashbrook and T. Starner, "Using GPS to Learn Significant Locations and Predict Movement Across Multiple Users," *Pers. Ubiquitous Comput*, vol. 7, no. 5, pp. 275–286, Oct. 2003.

[11] J. J.-C. Ying, W.-C. Lee, and V. S. Tseng, "Mining geographic-temporal-semantic patterns in trajectories for location prediction," *ACM Trans. Intell. Syst. Technol. TIST*, vol. 5, no. 1, p. 2, 2013.

[12] L. Bao and S. S. Intille, "Activity Recognition from User-Annotated Acceleration Data," in *Pervasive Computing*, 2004, pp. 1–17.

[13] T. Choudhury *et al.*, "The Mobile Sensing Platform: An Embedded Activity Recognition System," *IEEE Pervasive Comput.*, vol. 7, no. 2, pp. 32–41, Apr. 2008.

[14] E. M. Tapia, S. S. Intille, and K. Larson, "Activity Recognition in the Home Using Simple and Ubiquitous Sensors," in *Pervasive Computing*, 2004, pp. 158–175.

[15] D. J. Patterson, D. Fox, H. Kautz, and M. Philipose, "Fine-grained activity recognition by aggregating abstract object usage," in *Ninth IEEE International Symposium on Wearable Computers (ISWC'05)*, 2005, pp. 44–51.

[16] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity Recognition Using Cell Phone Accelerometers," *SIGKDD Explor Newsl*, vol. 12, no. 2, pp. 74–82, Mar. 2011.

[17] Y. Kim, F. C. Pereira, F. Zhao, A. Ghorpade, P. C. Zegras, and M. Ben-Akiva, "Activity Recognition for a Smartphone Based Travel Survey Based on Cross-User History Data," in *2014 22nd International Conference on Pattern Recognition*, 2014, pp. 432–437.

[18] L. Liao, D. Fox, and H. Kautz, "Location-based activity recognition," *Adv. Neural Inf. Process. Syst.*, vol. 18, p. 787, 2006.

[19] P. Nurmi and J. Koolwaaij, "Identifying meaningful locations," in *Mobile and Ubiquitous Systems-Workshops, 2006. 3rd Annual International Conference on*, 2006, pp. 1–8.

[20] G. Andrienko and N. Andrienko, *Extracting Patterns of Individual Movement Behaviour from a Massive Collection of Tracked Positions*. .

[21] OWASP, "Top 10 2017 - OWASP." [Online]. Available: https://www.owasp.org/index.php/Top_10_2017. [Accessed: 28-May-2017].

[22] A. M. Ortiz, E. Rios, W. Mallouli, E. Iturbe, and E. M. de Oca, "Self-protecting multi-cloud applications," in *2015 IEEE Conference on Communications and Network Security (CNS)*, 2015, pp. 643–647.

[23] A. D. Benedictis, V. Casola, M. Rak, and U. Villano, "Cloud Security: From Per-Provider to Per-Service Security SLAs," in *2016 International Conference on Intelligent Networking and Collaborative Systems (INCoS)*, 2016, pp. 469–474.

[24] S. O. Afolaranmi, L. E. G. Moctezuma, M. Rak, V. Casola, E. Rios, and J. L. M. Lastra, "Methodology to Obtain the Security Controls in Multi-Cloud Applications," in *Proceedings of the 6th International Conference on Cloud Computing and Services Science*, 2016, pp. 327–332.

[25] A. Kulkarni and M. G. McNally, "A Microsimulation of Daily Activity Patterns," *Cent. Act. Syst. Anal.*, Dec. 2000.

[26] J. Vieira, F. Dias, and A. Mota, "Comparison between artificial neural networks and neurofuzzy systems in modelling and control: a case study," *IFAC Proc. Vol.*, vol. 36, no. 12, pp. 249–255, Jul. 2003.

[27] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma, "Understanding Mobility Based on GPS Data," in *Proceedings of the 10th International Conference on Ubiquitous Computing*, New York, NY, USA, 2008, pp. 312–321.

[28] Y. Zheng, X. Xie, and W.-Y. Ma, "GeoLife: A Collaborative Social Networking Service among User, location and trajectory," *IEEE Database Eng. Bull.*, Jun. 2010.

[29] T. Bray, "The JavaScript Object Notation (JSON) Data Interchange Format." [Online]. Available: https://tools.ietf.org/html/rfc7159. [Accessed: 31-May-2017].

[30] Y. Ye, Y. Zheng, Y. Chen, J. Feng, and X. Xie, "Mining Individual Life Pattern Based on Location History," in *2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware*, 2009, pp. 1–10.

[31] C. L. Mallows and D. Richter, "Inequalities of Chebyshev Type Involving Conditional Expectations," *Ann. Math. Stat.*, vol. 40, no. 6, pp. 1922–1932, Dec. 1969.