SuperAD: Supervised Activity Discovery

Le T. Nguyen
Carnegie Mellon University
Moffett Field, CA, USA
le.nguyen@sv.cmu.edu

Ming Zeng
Carnegie Mellon University
Moffett Field, CA, USA
ming.zeng@sv.cmu.edu

Patrick Tague
Carnegie Mellon University
Moffett Field, CA, USA
patrick.tague@sv.cmu.edu

Joy Zhang
Carnegie Mellon University
Moffett Field, CA, USA
joy.zhang@sv.cmu.edu

Ubicomp/ISWC'15 Adjunct, September 07 - 11, 2015, Osaka, Japan

ACM 978-1-4503-3575-1/15/09$15.00
DOI: http://dx.doi.org/10.1145/2800835.2801617

Abstract
Activity recognition (AR) has become an essential component of many applications present in our everyday lives such as life-logging, fitness tracking, health and wellbeing monitoring. To build an AR system, one needs to first identify a set of activities of interest and collect labeled training data for these activities. However, activities of interest are not often known in advance. For example, a system designed to monitor a user's life style for potential diabetes risk needs to recognize all physical activities a user performs in her daily life. Given the large number of possible human activities, many of them cannot be foreseen during the model training time. In this work, we study the problem of discovering these unknown activities after the system is deployed by asking users to provide additional labels. Our goal is to discover all the unknown activities (i.e., obtain at least one label per class) while minimizing the amount of labels a user needs to provide. We propose SuperAD (Supervised Activity Discovery) approach, which combines active learning, semi-supervised learning and generative modeling to discover new unknown activities. We show that the proposed approach is especially effective when discovering activities with imbalance class distribution.

Author Keywords
Activity Recognition; Activity Discovery; Active Learning; Imbalance problem

ACM Classification Keywords

I.2.1 [Artificial Intelligence: Applications and Expert Systems]

Introduction

Applications such as life-logging and fitness tracking have become increasingly popular in today's self-quantification focused world. These applications rely on the ability of a system to automatically recognize human activities. In this work, we focus on systems recognizing activities from sensor readings of mobile and wearable devices.

To build an activity recognition (AR) system, one needs to first collect labeled training data for a set of activities of interest. However, activities of interest are often not known in advance. For example, a system designed to monitor a user's life style for potential diabetes risk needs to recognize all activities a user performs in his/her daily life. However, it is impractical if not infeasible to collect training data for all possible activities in advance, since 1) each user typically performs a different set of activities based on user's life styles and 2) there is a large number of possible activities, which are unknown to the system before the deployment (as illustrated in Figure 1).

A more practical solution is to build an AR system to recognize a small set of common activities, and after the deployment ask a user to provide additional labels to customize the AR system to the individual. The goal of collecting additional labeled data after the deployment is to discover new user activities unknown during the initial training phase. In this work, we study how to *discover new activities with minimal supervision*, i.e., we aim to obtain at least one labeled instance per activity class, while minimizing the total number of annotations required. This problem is further referred to as *supervised activity discovery*.

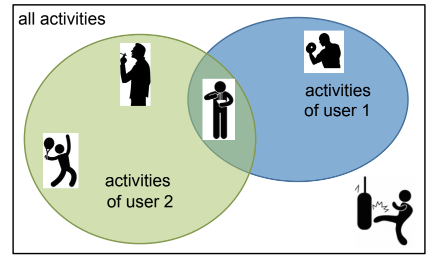


Figure 1: There is a large variety of activities each user performs in their everyday life. Discovering the presence of these activities is essential to many applications including life-logging or fitness tracking applications.

In the AR domain, the problem of minimizing supervision was studied in the context of active learning [12, 7, 1]. To improve the recognition results with minimal supervision, an active learning system asks a user to label only the most uncertain instances. Instances considered to be uncertain typically lie in between activity classes. Thus, labeling uncertain instances can help refine the decision boundaries between these classes.

Active learning has been shown to be effective especially when most of the activity classes are known [10]. However, in our scenario many user activity classes are not known prior to the deployment and need to be first discovered. Thus, using traditional active learning principles to label only uncertain instances does not necessarily help discovering the unknown activities. To illustrate the challenges of the discovery problem, we show an example of four activity classes in Figure 2. Activity class 1 and 2 are *discovered*, i.e., each of these activity classes we have at least one labeled instance (indicated by colors other than gray). Our goal is to discover remaining activity classes (activity classes 3 and 4). Using a traditional active learning model, a user is asked to label an instance close to the decision boundary. Thus, a system updated with a new label can improve its prediction performance on the instances close to the decision boundary, but it will fail to correctly recognize instances of the unknown activity classes (activity classes 3 and 4).

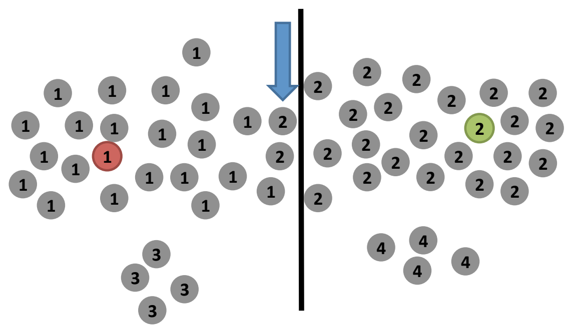


Figure 2: Given the labeled instances of the activity classes 1 and 2 (indicated by colors other than gray), an active learning system selects uncertain instances lying close to the decision boundary for labeling (indicated by the arrow). Such strategy typically leads to improving the prediction performance of instances lying close to the decision boundary, but does not necessarily help discovering the unknown activity classes (activity classes 3 and 4)

A naive solution for the supervised activity discovery problem is to use a random sampling method. This method is especially effective when activity classes are balanced, i.e., there is a equal number of instances per class. However, it is not clear, how imbalance activity distribution impacts the discovery performance.

In this work, we propose a **Supervised Activity Discovery (SuperAD)** approach, which builds on top of an active learning framework to discover new activities with minimal supervision. SuperAD uses semi-supervised learning and generative modeling to select instances that are likely "unknown" and ask users to label them. To the best of our knowledge, this work is the first to explore the problem of supervised discovery in the AR domain. Our contribution is summarized as follows:

- **Supervised activity discovery**: We propose an intelligent approach to discover activity classes with much less amount of human annotations compared to the state-of-the-art methods.
- **Imbalance activity distribution**: We study how the proposed approach performs on two public datasets with varying activity distributions.

Related Work

Supervised activity discovery is related to two concepts: unsupervised activity discovery and active learning.

Unsupervised activity discovery aims at discovering activities without any user annotations. This is done by finding sequences of sensor readings with similar patterns. These patterns can be discovered using unsupervised techniques such as frequent pattern mining, clustering or topic modeling [5, 14, 3, 13, 9]. Many of the unsupervised activity discovery approaches rely on frequent occurrences of the patterns to discover the activities. Thus, the more frequently a certain activity occurs, the more likely it will be discovered. However, this makes discovering infrequent activities a challenging task, especially in cases of imbalanced activity distribution.

In addition, the output of the unsupervised activity discovery does not have a semantic meaning. The results of the discovery can be interpreted only if the actual ground truth is known. This is possible when evaluating the system before the deployment, but infeasible once the system is deployed in the wild and no ground truth is available for interpreting the discovery results. On the other hand, unsupervised discovery of activities is a highly valuable capability, since it can uncover inherent patterns of sensor readings. In future work, we will study how to combine such capability with the proposed supervised discovery approach to further reduce the amount of requested labels.

Active learning has been studied in many previous works as a means for improving the AR performance with a minimal supervision [12, 7, 1]. It leverages the unlabeled data to identify uncertain/informative instances for labeling. Active learning performs well when all relevant activity classes are known, since it helps the system to refine the decision boundaries between the activities. However, in our scenario only a small amount of activities are known in advance and many activities need to be discovered first.

The goals of active learning and supervised activity discovery are divergent and often contradictory. While active learning aims at finding uncertain instances, which lie close to the decision boundary, active discovery aims at finding instances of new classes, which typically lie far away from the existing activity classes. Thus, active learning does not necessarily lead to fast discovery of new activities.

In this work, we propose an approach which extends active learning to address the supervised activity discovery problem. In the following, we first describe the traditional active learning process. In the next section, we explain how our proposed approach extends active learning to discover new activities.

The active learning process can be divided in four steps as depicted in Figure 3. First, it is assumed that we have a small amount of labeled data and a large amount of unlabeled data. The labeled data is used to build an initial AR model. This model is then iteratively updated with new labels requested from the user. To select instances for labeling, the system performs the following steps. It uses the current AR model to predict the probability distribution $p(y|x)$ for each unlabeled instance x:

$$p(y|x) = [p(y_1|x), p(y_2|x), \cdots, p(y_N|x)] \tag{1}$$

where N is the number of known activities.The system then selects a candidate for labeling x^* based on one of the strategies described below. Once a user labels the candidate instance, the new instance is added into the labeled pool and the AR model is retrained.

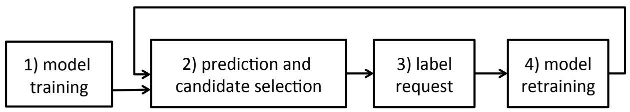


Figure 3: Active learning process involves four steps: First an initial AR model is trained, which is then iteratively updated through the cycle of prediction, candidate selection, labeling and retraining.

There are three common strategies used for selecting the candidate for labeling x^*:

1. *Entropy-based approach*:

$$x^* = \arg\min_x \sum P(y|x) log P(y|x) \tag{2}$$

2. *Margin-based approach*:

$$x^* = \arg\min_x [P(\hat{y_1}|x) - P(\hat{y_2}|x)] \tag{3}$$

where $\hat{y_1}$ and $\hat{y_2}$ are the most probable and the second most probable class of x.

3. *Confidence-based approach*:

$$x^* = \arg\min_x P(\hat{y}|x) \quad (4)$$

where $\hat{y}$ is the most probable class of x.

These strategies differ from each other in terms of how much information from the predicted probability distribution is used for selecting the next candidate for labeling. The entropy-based strategy uses all values in the probability distribution, whereas the confidence-based and margin-based approaches use only the highest values or the two highest values in the predicted probability distribution.

SuperAD: Supervised Activity Discovery

SuperAD is built on top of a pool-based active learning model, i.e., it assumes that a user can label any instance that is presented to him/her. This assumption is commonly made in the existing active learning work in the AR domain [12, 7, 1]. This assumption is valid in scenarios when besides carrying a mobile sensing device such as a smartphone or a smartwatch, users also wears a wearable camera in their daily life in order to aid the annotation process [4]. At the end of each day, SuperAD 1) analyzes the sensor data collected during the day, 2) identifies activity candidates for labeling and 3) presents a photo or a video snippet captured at the time when an activity of interest was performed to help the user with data annotation. A user needs to wear the wearable camera only for a certain number of days to help the system discover most of the user activities. In future work, we will study how to extend SuperAD to stream-based active learning conditions [11], where the user is asked in real-time to annotate activities, thus eliminating the need of wearable cameras.

Our work is based on the work of Pelleg et al. [10] with extensions to address issues encountered in real world AR applications. In the following, we describe the SuperAD algorithm and discuss its differences from the existing work.

Algorithm: Initially we have a small set of labeled instances and a large amount of unlabeled data.

1. Given the labeled set, we train a supervised classifier.
2. This classifier is then used to predict a "temporary" label for each unlabeled instance.
3. The unlabeled instances are grouped together based on the "temporary" label and for each label we train one Gaussian Mixture Model (GMM).
4. We use the GMMs to estimate the likelihood of each unlabeled instance being generated by its corresponding GMM.
5. Instance with a lowest likelihood is selected as a candidate for labeling.
6. After obtaining the label, we go back to step 1.

Initially, it is assumed that a small set of labeled instances is available to train the initial classification model. In case no labeled data are available, we use k-means ($k = 2$) to cluster the unlabeled instances and for each cluster select an instance closest to its centroid for the initial label request.

In step 1, we build a classifier corresponding to the initial AR model, which can be either a discriminative (e.g., SVM, Random Forrest, etc.) or generative model (e.g., Naive Bayes). This model is iteratively updated using the active learning process.

In step 2 and 3, we use a semi-supervised approach to make use of both the small amount of labeled data and the large amount of unlabeled data. First, the initial AR model used to predict the labels of the unlabeled instances. Thus, each instance obtains an “temporary label”. This labels are used to train the GMM models, specifically, for each discovered activity class we train one GMM model from all instances, which are predicted to belong to this activity class (as shown in Figure 4).

In step 4 and 5, we aim to select the candidate for labeling using the trained GMMs. First, for each instance we use the corresponding GMM to estimate the likelihood that the GMM generated this instance. Then, we select an instance with the lowest likelihood. The key idea is to identify instances, which are not least likely to be from their predicted activity class and thus we hypothesize that they are more likely to be undiscovered activities so we check with the user to get them labelled.

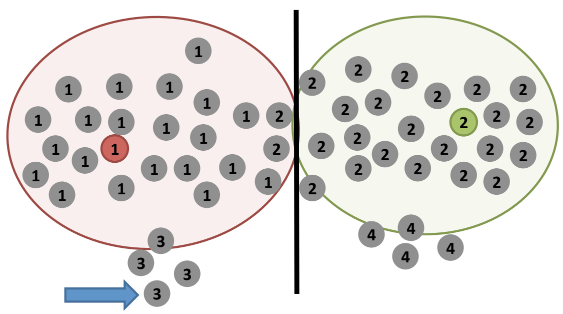


Figure 4: SuperAD learns for each discovered activity class one GMM (Gaussian Mixture Model) using all instances predicted to be of this activity class. An instance with the lowest likelihood of being generated by its GMM is selected as the candidate for labeling (indicated by the arrow).

Discussion: Pelleg et al. [10] use Gaussian Bayes classifier, which models each class using one single Gaussian. This however assumes that instance follow an Gaussian distribution, which is not necessarily the case in AR. Due to high variability of activities, instances of the same class might appear in two disjoint subspaces in the feature space. SuperAD uses a GMM instead, which can theoretically approximate any distribution (including disjoint distributions). Moreover, Pelleg et al. [10] select a batch of candidates for labeling in each iteration, which might lead to having users label many similar and thus redundant instances per iteration. In this work, in each iteration we select only one instance for labeling to reduce the redundancy.

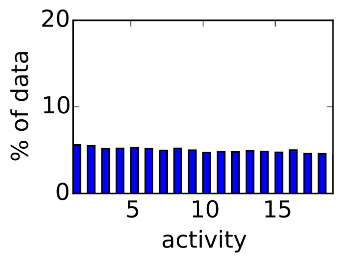
Compared to existing active learning approaches presented in the previous section, SuperAD leverages a two-stage strategy to make use of the large pool of the unlabeled data. The first stage of SuperAD is similar to the traditional active learning approach, while in the second stage, SuperAD converts the unlabeled data into a temporary labeled data to learn a generative GMM model. This semi-supervised strategy of SuperAD makes use of the large pool of information ignored by the traditional active learning approaches.

Evaluation

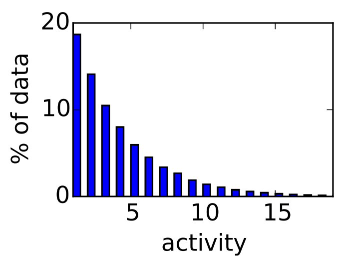
In the following, we conduct two sets of experiments to evaluate the proposed approach with respect to the speed of activity discovery and the imbalanceness of the activity distribution. In the first set of experiments we use a balanced set of activities collected in a controlled setting and create various imbalanced scenarios to study the effectiveness of the proposed approach. In the second set of experiment we use a dataset collected from a real-world application, to study the activity discovery in natural imbalanced setting.

Discovering Activities with Imbalanceness

Data: In the first experiment, we use the Daily and Sport activity dataset [2] containing 19 activity classes performed



(a) $p = 0.01$



(b) $p = 0.25$

Figure 5: An imbalanced dataset is generated by using a subsampling method based on the geometric distribution with a parameter p. High p corresponds to high imbalanceness.

by 8 users. Each user performs each activity 60 times, each time for a duration of 5 seconds. Each 5-second segment of sensor data is used to extract a feature vector (further referred to as an instance) composed of statistical features including mean, standard deviation, minimum, maximum, energy and correlation between sensor axes of individual sensors [6]. Note that this dataset is a balanced dataset, i.e., each class has exactly the same amount of instances.

Imbalanceness: To study the effectiveness of the proposed approach, we use a subsampling method based on a geometric distribution to obtain a dataset with a certain imbalanceness property. First, we define p as an imbalanceness score, which corresponds to the parameter of the geometric distribution. With p close to zero (Figure 5a), we obtain a balance dataset with each activity class having a similar amount of instances. On the other hand, high p indicates a high imbalanceness (Figure 5b), i.e., a resulting dataset will be dominated by a few frequent activities.

To generate a dataset with an imbalance score p and size N, we use the following strategy: In each iteration i: 1) we randomly select one activity class and 2) sample m_i instances of this class from the original dataset:

$$m_i = max(5, p(1-p)^{i-1} \times N) \tag{5}$$

We repeat this iterative process until all activities are selected and sampled. This methods ensures the order of selected activities are random, i.e., frequent activities are chosen from the set of all activity classes. Moreover, the max function ensures that each activity class contains at least 5 instances. In the following, we set N to 2000, i.e., a new dataset is created by randomly selecting approximately 2000 instances from the total 9120 instances.

In the following experiment, we compare the performance of SuperAD with the existing active learning approaches described in the previous section and with random sampling. As the base classifier for all approaches we evaluated both SVM and Random Forest. Due to their comparable results, we show in the following only the results of the Random Forest classifier. For SuperAD we empirically set the number of Gaussian components in GMM to 3. High number of Gaussians components typically leads to overfitting whereas low number of Gaussians would reduce its capability to capture the complex data distributions. In the future work, we will further study the proposed approach using different set of configurations.

Speed of Activity Discovery: First we evaluate how many instances a user needs to label to discover a certain amount of activities. Figure 6 shows the results for $p = 0.25$ for one single run. From the results we can observe that SuperAD outperforms the other approaches in terms of the number of classes discovered for any number of labeled instance.

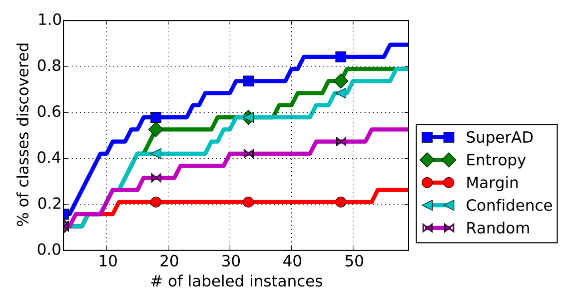


Figure 6: Daily and Sport dataset: The percentage of discovered activities given the number of instances labeled by a user (for the imbalance score $p = 0.25$).

The entropy-based approach selects instances equally far away from all discovered classes (Equation 2). Thus, it has a potential to discover new activities. However, equally far from all classes means either lying close to the boundary between the classes or lying far away for all classes. The former case corresponds to refining decision boundaries, while the latter case corresponds to discovering new activities. Thus, the entropy-based approach alternates between the discovery and boundary refinement modes. Since SuperAD focuses only on the activity discovery, it achieves better results than the entropy-based approach.

The margin-based approach achieves the worst results, since it is optimized only towards refining decision boundaries. As shown in Equation 3, in each iteration a candidate selected for labeling if it lies between two classes. Thus, the margin-based approach aims at obtaining labels for instances, which can refine the already known approximate boundaries instead of focusing discovering new activities.

The confidence-based approach is the most similar to the SuperAD, since it aims at selecting predictions with lowest confidence (Equation 4). However, in each iteration the confidence-based approach uses only the limited labeled to build the model for estimating the confidence. SuperAD outperforms the confidence-based learning approach, since it leverages the additional the unlabeled data in its two stage semi-supervised learning approach.

The random sampling approach achieves a relatively low percentage of discovered classes. This is caused by the fact that the activity distribution of $p = 0.25$ is skewed, i.e., a few activities occur frequently, while a large number of other activities are highly infrequent. Thus, selecting instances randomly is not an effective discovery strategy, since the most of the time the instance of the most frequent activity class is selected.

Speed of Activity Discovery vs. Imbalanceness: In the following, we show how the imbalanceness of activity distribution (p) impacts the discovery speed. To compare the results across different values of p, we use the normalized Area Under the Curve (AUC) measure:

$$AUC = \frac{1}{N} \sum_{n=1} c_n \tag{6}$$

where c_i corresponds to the number of activities discovered for n labeled instances and N corresponds to the number of active learning iterations. This measure basically computes the area under the curve shown in Figure 6 and normalizes it based on the total number of active learning iteration. The higher AUC the faster the activities are discovered. In the following we report the AUC averaged across 10 runs, since the process of generating dataset for each experiment involves a random process of selecting instances.

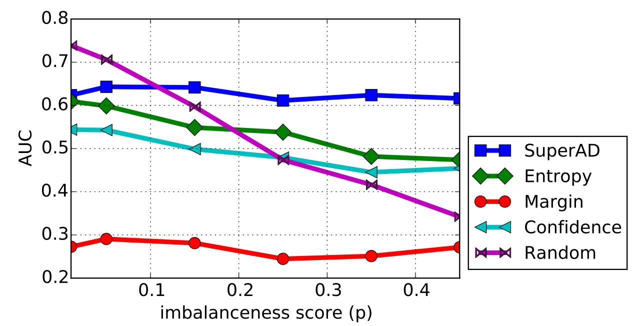


Figure 7: TU Darmstadt dataset: The performance of SuperAD remains the same while other performance of other approach degrades with the increasing imbalanceness.

Figure 7 shows AUC for different values of p. From the results we can observe that for balanced datasets (p close to 0), the random sampling performs the best, since it can effectively discover the new activities with a uniform distribution. However, the performance of random sampling degrades with the increase of the imbalanceness. This is

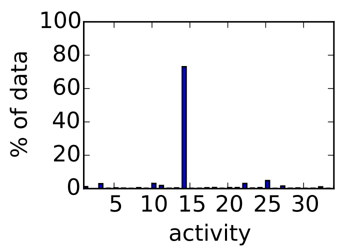


Figure 8: TU Darmstadt dataset is highly imbalanced.

caused by the fact, that in imbalanced datasets, there are only a few frequent activities, which dominates this dataset. Thus, the random sampling is likely to sample instances of frequent activities while ignoring other infrequent activities.

The performance of SuperAD remains constant for all p values. However, the performance of entropy-based and confidence-based approaches decreases with the imbalance, as these activities are less effective at discovering rare activities. The margin-based approach performs the worst as discussed in the previous subsection.

Discovering Activities in Real-World

Data: To evaluate the proposed approach in the real-world setting, we use the TU Darmstadt dataset [5], containing data collected by one user in a natural setting for 7 days containing 33 activity classes. As shown in Figure 8, the class distribution is highly imbalanced, i.e., one activity ("sitting and working") dominates the whole dataset in terms of the amount of time this activity occurs in the user's daily life.

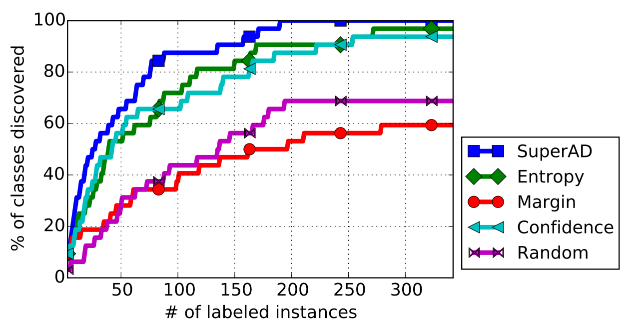


Figure 9: Compared with the second best solution (entropy-based approach), SuperAD can discover 85% of the activity classes while needing only a half of the amount of user labels.

Speed of Activity Discovery: As shown in Figure 9, SuperAD consistently outperforms the other approaches. Moreover, to discover 85% of the activity classes, the user needs to label 77 instances. To achieve comparable results, the second best approach (entropy-based approach) needs 150 labeled instances, i.e., almost double the amount of labeled instances is needed. Regarding other approaches, we can observe similar trends as in the Figure 6

Conclusion and Future Work

In this work, we study the supervised activity discovery problem and identify its unique challenges. We propose SuperAD aiming at discovering activities while minimizing the amount of user annotations. We show that SuperAD is especially effective in scenarios with imbalanced activity distribution, which commonly occurs in real-world setting.

In this work, we mainly focus on the speed of discovery. Obviously, a good discovery algorithm would require a user to label as little amount of data as possible. This, however, results in obtaining only a small amount of training data for each activity class. As it was shown in the prior work, AR models learned with a small amount of training data may result in poor AR performance [8]. Thus, in the future work, we will study how to combine SuperAD with existing techniques aimed at learning with limited training data. Moreover, we will study how to convert SuperAD to a stream-based active learning setting to increase the practicality of the proposed system.

Acknowledgement

This research is supported in part by the National Science Foundation under the award 1346066: SCH: INT: Collaborative Research: FITTLE+: Theory and Models for Smartphone Ecological Momentary Intervention.

References

[1] Hande Alemdar, Tim LM van Kasteren, and Cem Ersoy. 2011. Using active learning to allow activity recognition on a large scale. In *Ambient Intelligence*. Springer, 105–114.

[2] Kerem Altun, Billur Barshan, and Orkun Tunçel. 2010. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognition* 43, 10 (2010), 3605–3620.

[3] Diane J Cook, Narayanan C Krishnan, and Parisa Rashidi. 2013. Activity discovery and activity recognition: A new partnership. *Cybernetics, IEEE Transactions on* 43, 3 (2013), 820–828.

[4] Katherine Ellis, Suneeta Godbole, Jacqueline Kerr, and Gert Lanckriet. 2014. Multi-Sensor physical activity recognition in free-living. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM, 431–440.

[5] Tâm Huynh, Mario Fritz, and Bernt Schiele. 2008. Discovery of activity patterns using topic models. In *Proceedings of the 10th international conference on Ubiquitous computing*. ACM, 10–19.

[6] Oscar D Lara and Miguel A Labrador. 2013. A survey on human activity recognition using wearable sensors. *Communications Surveys & Tutorials, IEEE* 15, 3 (2013), 1192–1209.

[7] Rong Liu, Ting Chen, and Lu Huang. 2010. Research on human activity recognition based on active learning. In *Machine Learning and Cybernetics (ICMLC), 2010 International Conference on*, Vol. 1. IEEE, 285–290.

[8] Le T. Nguyen, Ming Zeng, Patrick Tague, and Joy Zhang. 2015. Recognizing New Activities with Limited Training Data. In *International Symposium on Wearable Computers (ISWC)*. IEEE.

[9] Le T. Nguyen and Joy Zhang. 2014. Unsupervised work knowledge mining through mobility and physical activity sensing. In *International Conference on Mobile Computing, Applications and Services (MobiCASE)*. IEEE, 30–39.

[10] Dan Pelleg and Andrew W Moore. 2004. Active learning for anomaly and rare-category detection. In *Advances in Neural Information Processing Systems*. 1073–1080.

[11] Burr Settles. 2010. Active learning literature survey. *University of Wisconsin, Madison* 52, 55-66 (2010), 11.

[12] Maja Stikic, Kristof Van Laerhoven, and Bernt Schiele. 2008. Exploring semi-supervised and active learning for activity recognition. In *Wearable Computers, 2008. ISWC 2008. 12th IEEE International Symposium on*. IEEE, 81–88.

[13] Feng-Tso Sun, Yi-Ting Yeh, Heng-Tze Cheng, Cynthia Kuo, and Martin Griss. 2014. Nonparametric discovery of human routines from sensor data. In *Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on*. IEEE, 11–19.

[14] Alireza Vahdatpour, Navid Amini, and Majid Sarrafzadeh. 2009. Toward Unsupervised Activity Discovery Using Multi-Dimensional Motif Detection in Time Series.. In *IJCAI*, Vol. 9. 1261–1266.