Exploring Semi-Supervised and Active Learning for Activity Recognition

Maja Stikic
Fraunhofer IGD, Germany
stikic@mis.tu-darmstadt.de

Kristof Van Laerhoven
TU Darmstadt, Germany
kristof@mis.tu-darmstadt.de

Bernt Schiele
TU Darmstadt, Germany
schiele@mis.tu-darmstadt.de

Abstract

In recent years research on human activity recognition using wearable sensors has enabled to achieve impressive results on real-world data. However, the most successful activity recognition algorithms require substantial amounts of labeled training data. The generation of this data is not only tedious and error prone but also limits the applicability and scalability of today's approaches. This paper explores and systematically analyzes two different techniques to significantly reduce the required amount of labeled training data. The first technique is based on semi-supervised learning and uses self-training and co-training. The second technique is inspired by active learning. In this approach the system actively asks which data the user should label. With both techniques, the required amount of training data can be reduced significantly while obtaining similar and sometimes even better performance than standard supervised techniques. The experiments are conducted using one of the largest and richest currently available datasets.

1. Introduction

Activity recognition is an important and active research area of wearable computing due to its potential to enable novel context-aware applications for elderly care, education, sports, and entertainment. Different types of sensors have been proposed for this purpose ranging from motion sensors such as accelerometers [2, 16] over RFID tag readers [25] and switch sensors [19] to combinations of different sensor modalities [17, 22, 23].

Most approaches for human activity recognition are based on state-of-the-art machine learning techniques. An important difference between approaches is whether they rely on supervised training or whether they enable the use of unsupervised learning techniques. A wide range of supervised classification approaches have been applied for the recognition of both "simple" activities (e.g., sitting, standing, lying, and walking [10, 11]) and more complex or composed activities (e.g., Activities of Daily Living [2, 19, 25], wood workshop activities [23], and maintenance tasks [17]). These techniques can be categorized as either generative algorithms that model class-conditional distributions [2, 23] or discriminative algorithms that focus on learning the class decision boundaries [11, 16].

The main drawback of supervised methods is the necessity of a significant amount of labeled data for learning activity models. Labeling data for activity recognition systems is a challenging problem for at least two reasons. First, most of the annotation techniques are time-consuming and error-prone. And second, to obtain reliable annotations one has essentially two choices. Either one may rely on invasive sensors such as cameras and microphones [12] which are often not acceptable due to privacy reasons. Or, one uses annotation techniques such as experience sampling [19] which is tedious or disrupting for users in particular for annotation of short term activities.

Another line of research avoids the labeling efforts by unsupervised discovery of structure in activity data [6, 9, 14]. Also, it is possible to define prior models for activities by manually specifying common sense features of activities [22] or automatically extracting this information from the web [25]. However, while the learned structure results in interesting representations of the data one still requires at least a few labels to achieve reliable classification results.

In many practical classification problems, data labeling is expensive, but a large number of unlabeled data can be easily obtained. For this reason, semi-supervised learning has been proposed as an alternative in machine learning research [4]. The ultimate goal of semi-supervised learning is to combine the advantages of supervised and unsupervised approaches by learning from both labeled and unlabeled data. Since many human activities of interest are performed on a daily basis, it is relatively easy to produce large quantities of unlabeled activity data. Thus, semi-supervised learning naturally lends itself to activity recognition.

The primary goal of this paper is to explore and compare two different types of techniques that require far less labeled training data than traditional supervised techniques. First, we apply and analyze the merits of two of the most fundamental semi-supervised learning techniques, namely

978-1-4244-2638-6/08/$25.00 ©2008 IEEE

self-training and co-training. And second, we also explore another way to reduce the required amount of labeled training data. This second approach is based on active learning [15] with the explicit goal to focus labeling effort on the most profitable, e.g. informative, instances of activities. There exists relatively little work [5, 13, 18] exploring semi-supervised techniques for human activity recognition. However, these approaches do neither address nor analyze the potential of active learning for the recognition of physical activities. Additionally, the evaluation of the proposed approaches was performed on relatively simplistic datasets consisting mostly of activities such as sitting, standing, walking, and running. On the other hand, the focus of active learning approaches [1, 8] is on the recognition of user's desktop activities for predicting interruptibility of a user. We make a first step towards active learning for physical activity recognition.

The main contributions of the paper are as follows. First, we present a comparative evaluation of the applicability of self-training [4] and co-training [3] for data from motion sensors. Unlike in [5], where an ensemble method based on one set of features has been proposed, we show that it is possible to apply co-training for recognition of activities when using two independent sources of information, namely on-body accelerometers and infra-red motion sensors. Second, we suggest two functions to actively probe users for labels that enable active learning. The wrapper nature of the proposed semi-supervised algorithms and active sampling functions makes them independent of both classifiers and sensor modalities being used. Additionally, their low computational costs are very beneficial for enabling wearable computing scenarios. Third, we enhance the efficiency of the proposed activity recognition system by utilizing a multi-class boosting procedure, namely joint boosting [21]. Additionally, the typical researchers' bias on the evaluation is avoided by using a publicly available dataset [12] that was neither recorded nor annotated by the authors of the paper. By using only a limited amount of labeled training data, we achieve performance comparable to and sometimes even better than fully supervised learning approaches on a challenging and realistic dataset.

The rest of the paper is organized as follows. In Section 2 we introduce the dataset and sensors used in the experiment as well as our evaluation procedure. Section 3 presents the initial supervised analysis of the dataset followed by our semi-supervised and active learning approaches in Section 4 and Section 5, respectively. Finally, in Section 6 we summarize our results and give an outlook on future work.

2. Experimental Setup

In this section, we present the goals of our experiment, motivate the choice of the used dataset and describe the evaluation procedure. In the field of activity recognition, the state-of-the-art has advanced significantly in recent years and a wide range of sophisticated approaches and sensors has been developed. An important drawback of the majority of the current activity recognition systems is the lack of a standardized evaluation procedure that would enable a unified way of comparison of different approaches. Thus, in this work we follow a different approach.

We obtained access to the subset of the PLCouple1 dataset [12] recorded at the PlaceLab [7], a highly instrumented home environment, where a couple moved in and lived there for 10 weeks, continuing as normal a routine as possible. An audio-visual recording system was used for capturing the ground truth and an expert annotated 104 hours of the male's activities, comprising data collected on 15 separate days. In our experiment we use a publicly available subset of 68 hours of annotated data collected on 9 separate days. Despite a substantial amount of data collected and annotated, there is still a lack of data for many fine-grained activities, which led to 9 activities to be studied in [12]. Here we focus on the same set of activities: *actively watching tv or movies*, *dishwashing*, *eating*, *grooming*, *hygiene*, *meal preparation*, *reading paper/book/magazine*, *using computer*, and *using phone*.

In our experiment, we use the data from two different types of motion sensors [20], namely body-worn accelerometers and infra-red sensors. In [12], these two sensor modalities outperformed other sensors (i.e., RFID and environmental built-in sensors). The male subject wore 3 3D accelerometers on the dominant wrist, the dominant hip, and the non-dominant thigh. Ten infra-red sensors were installed around the apartment to detect motion in each room.

Unlike in [12] where the mean value of the acceleration signal and binary occurrences of the infra-red readings were used as features, we extract the following features to exploit the full richness of information in the data: 1) From the raw acceleration signal we compute *mean*, *variance*, *energy*, *spectral entropy*, *area under curve*, *pairwise correlation between the three axes*, and the first ten *FFT coefficients*, which sums up to 48 features per acceleration sensor channel. 2) For each of the ten infra-red sensors we calculate the number of their activations as features. As in [12], each feature is computed over a sliding window of 30 seconds shifted in increments of 15 seconds. We experimented with different window lengths as well, but that did not significantly change performance.

As suggested in [12], we use 9-fold leave-one-day-out cross validation on the data to avoid over-fitting. In each cross validation round of supervised learning, we train the algorithms on 8 days of data. In case of semi-supervised and active learning, only a subset of 2 days of data is used as an initial labeled training set. The algorithms are always tested on the left out day's data.

3. Supervised Approach

As we use the publicly available subset of the PLCouple1 dataset, we first reproduce the experiments from [12] based on two supervised machine learning algorithms (i.e., naive Bayes and decision tree). Additionally, we compare their performance to the joint boosting classifier [21]. These results are used as a baseline for comparison with semi-supervised and active learning approaches in Section 4 and Section 5, respectively.

Naive Bayes is a simple yet effective generative classifier that has been used in the field of activity recognition (e.g., [10], [16]). Even though it assumes that the components of a feature vector are independent of each other, it often outperforms more sophisticated classifiers. For our experiments we use the unimodal Gaussian model for acceleration data. For infra-red data we use the multinomial model when using the number of the activations as a feature and the multi-variate Bernoulli model for binary features.

Decision tree learning is based on inductive inference and it has also been successfully used for activity recognition (e.g., [2], [16]). We use the C4.5 variant of a decision tree algorithm found in the Weka Machine Learning Algorithms Toolkit [24].

Joint boosting [21] is a multi-class variant of traditional boosting approaches in which multiple weak learners are combined into a single strong classifier. Each weak learner is a decision or regression stump on a single component of a feature vector. Joint boosting is especially appealing because it finds the features that can be shared across the classes, which results in a faster classifier that needs less features than standard approaches. At each boosting round different subsets of classes are examined for fitting a weak learner to distinguish that subset of classes from the other classes. The subset that maximally reduces the error on the weighted training set for all the classes is chosen. The best weak learner is then shared among the classes in that subset. More details about the algorithm can be found in [21].

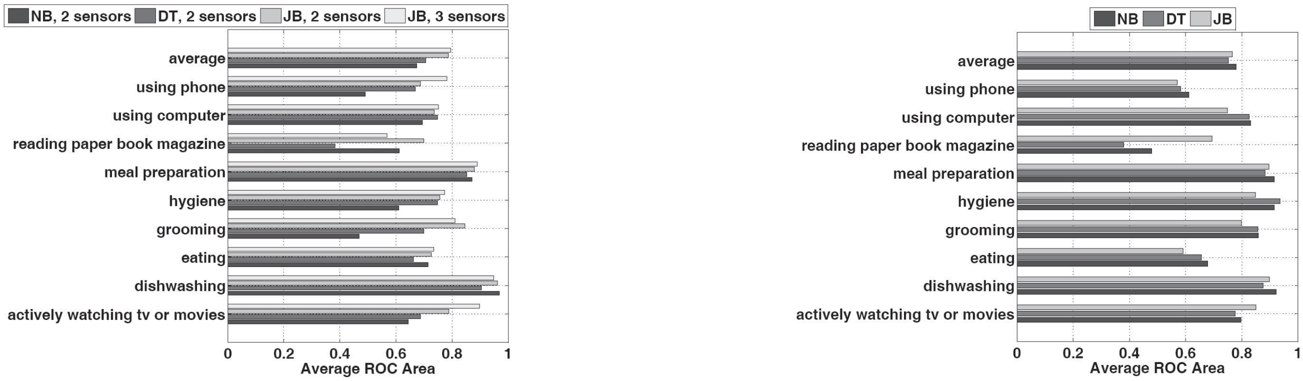
As the dataset contains partly overlapping activities that are not mutually exclusive, we use, as in [12], the area under the Receiver Operating Characteristic (ROC) curve, averaged over 9 cross validation rounds, as a figure of merit. The ROC curve plots the true positive rate vs. the false positive rate, and it provides an overall measure of goodness at all possible thresholds of a classifier. For naive Bayes and decision tree we apply a “one vs. the rest of the world” approach, as in [12], by using a binary classifier for each activity. The main drawback of that approach is that it does not deal well with highly unbalanced datasets. The overall duration of activities in the dataset strongly varies among the activities, reflecting the natural distribution of activities in real life. Thus, the balancing of the training set had to be done, as in [12], by uniformly sampling the examples from the “rest of the world” class to match the number of examples in the class, i.e., activity of interest. Interestingly, joint boosting, being a multi-class classifier, lends itself to joint training on all classes by finding features that can be shared across the classes. As a consequence, it is even able to deal properly with multi-label data of overlapping activities (i.e., activities that were performed in parallel which resulted in multiple labels for a single sample). We transformed multi-label samples to single-label samples as follows: Each multi-label sample consisting of n labels is replicated n times, and the i-th copy is assigned the i-th label. During the classification we accept all classes with classification scores higher than a certain threshold.

In [12] movement data measured by two accelerometers, worn on the dominant wrist and on the dominant hip, were used. We performed the experiments with both two and three accelerometers since the addition of sensors often improves recognition performance.

3.1. Results

In the following we report the recognition results based on the supervised algorithms. We experimented with both binary features and the number of the activations of infra-red sensors. Due to space constraints, we only report the best results per classifier, i.e. performance of naive Bayes and decision tree for binary features and performance of joint boosting when using the number of activations as a feature. We perform the experiments with different numbers of joint boosting rounds. The best performance is achieved after 50 iterations for acceleration data and after 10 iterations for infra-red data. Since the acceleration feature vector has 144 components it requires more boosting rounds to find the best features to be shared among the activities. The infra-red feature vector has only 10 components and weak learners from additional rounds could not contribute to the better performance.

Figures 1(a) and 1(b) show results per activity and average recognition performance for acceleration and infra-red sensors, respectively. A few trends stand out. First, one can observe that joint boosting yields better results for 7 out of 9 activities when using acceleration data. On average, joint boosting improves the results by 11.3% compared to naive Bayes and by 8.2% compared to the decision tree classifier. Second, the addition of the third accelerometer does not improve the results significantly, presumably because the placement of the sensor at the non-dominant thigh is not discriminative for the majority of the activities studied. Third, naive Bayes on average performs slightly better for infra-red sensors. As stated in [12], the presence of a second subject in the apartment whose activities were not annotated definitely introduced noise in the infra-red sensor data. Thus, naive Bayes, as a generative model, is able to



(a) Per-activity and average results for acceleration data

(b) Per-activity and average results for infra-red data

Figure 1. Leave-one-day-out cross validation results for supervised classifiers (naive Bayes - NB, decision tree - DT, and joint boosting - JB)

Sensor	Accuracy	Average ROC Area
Acceleration	53.6%	79.3%
Infra-red	41.6%	68.6%

Table 1. Leave-one-day-out cross validation results for joint boosting classifier on single-label subset of the dataset

deal better with the noisy data compared to the joint boosting and decision tree classifiers. Even though, we use only the publicly available subset of the PLCouple1 dataset, the decision tree results are nearly the same as reported in [12].

As previously mentioned, the dataset contains a certain amount of overlapping activities. The multi-label data constitutes about 10% of the whole dataset. Table 1 summarizes the classification results of the joint boosting classifier when leaving out the multi-label part of the dataset. The results are consistent with the multi-label case (i.e., joint boosting again performs better on acceleration data). Additionally, the table shows accuracy of the classification, i.e., the number of true positives divided by the number of all samples to be classified. One can observe that accuracy is relatively low (53.6% for acceleration data and 41.6% for infra-red data), but that should be seen in the light of realism of the used dataset which additionally includes many other activities that were considered as an unknown class during the classification procedure. In order to thoroughly explore the potential of semi-supervised and active learning in activity recognition we decided to use a clean dataset (i.e., without multi-label samples) in the remainder of the paper. The results in Table 1 are used as a baseline for comparison with semi-supervised and active learning approaches. As a figure of merit we use accuracy, which we consider more intuitive and which is more often used than the area under the ROC curve in the field of activity recognition.

4. Semi-Supervised Approaches

In this section we introduce the two semi-supervised approaches, self-training and co-training, which we use in our experiments for learning from both labeled and unlabeled training data. Typically, in semi-supervised settings, it is assumed that in addition to the small set of labeled training data there is also a substantial amount of unlabeled training data available. This allows reducing the effort of supervision to a minimum, while still preserving competitive recognition performance.

Self-training [4] is a wrapper-algorithm that repeatedly uses a supervised learning method in the following manner. A supervised classifier is first trained with a small amount of labeled data. The classifier is then used to classify the unlabeled data. In each iteration, a part of the unlabeled data is labeled according to a current decision function. Typically, the most confident predictions are added to the labeled training set. The classifier is then re-trained and the self-training procedure is repeated.

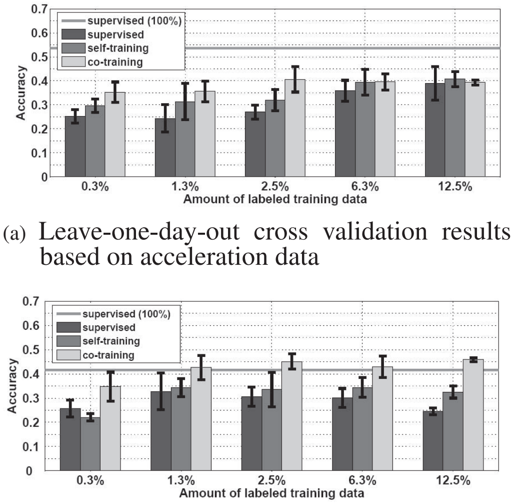
Co-training [3] follows the iterative training procedure of self-training. At the same time, it aims to improve self-training by augmenting the training process with an additional source of information. Thus, we initially use acceleration and infra-red feature sets for training two separate classifiers. Classifiers then teach one another by augmenting each other's training sets with their most confident predictions. The classifiers are then re-trained with the refined labeled training sets and the process is iteratively repeated. Co-training is based on the two assumptions that are fulfilled in our multi-sensor approach. First, it assumes that features can be split into two disjoint sets that are sufficient for learning in the supervised setting so that one can trust the predictions based on both sets. Second, the two sets of features need to be independent given the class, so that one classifier's high confident data points are independent and identically distributed samples for the other classifier.

In [5] it has been argued that co-training is not applicable to activity recognition due to the strong independence assumption. In this paper, we show that co-training is an excellent method for activity recognition approaches that aim at improving recognition results by fusing different sensor modalities. In the following experiments we use acceleration and infra-red data for co-training and compare its performance with self-training. Since joint boosting shows superiority compared to the naive Bayes and decision tree classifiers, we use it as the supervised part of the self-training and co-training procedure.

The experiments are designed to investigate the trade-off between labeling efforts and recognition performance. The goal of the experiments is to decrease the amount of necessary labeled training data to a minimum. For that purpose, we use the following evaluation procedure. Leave-one-day-out cross validation is again performed by using one day of data for testing and the remaining eight days of data for training. The distribution of activities varies significantly for different days. Since we want to find the lower boundary for the size of labeled training data we use a minimum amount of data to have at least one representative for each of the activities of interest. In case of the used PLCouple1 dataset, that means that we can use six days of data as unlabeled training set and the remaining two days of data as an initial set for subsampling to get the reduced set of labeled training data. The experiments consist of five different configurations in which we gradually decrease the amount of labeled training data from 12.5%, over 6.3%, 2.5%, 1.3% to 0.3% of 8 days of training data.[1] In order to thoroughly analyze the classifiers' performance we perform multiple random subsampling rounds. The reported results are averaged over 9 cross-validation and 5 random subsampling rounds.

4.1. Results

An important parameter of self-training and co-training algorithms is the number of iterations. By conducting experiments with different numbers of iterations we observed that by performing more than 100 iterations the newly labeled samples do not bring any additional discriminative information, and at a certain point the labeling accuracy even starts to decrease. For comparison of self-training and co-training, in the following, we report on the average recognition accuracy achieved after 100 iterations.



(a) Leave-one-day-out cross validation results based on acceleration data

(b) Leave-one-day-out cross validation results based on infra-red data

Figure 2. Comparative performance of self-training, co-training and supervised learning for different amounts of labeled training data

We also observed that for our multi-class problem it is crucial to maintain the underlying distribution of activities. In each iteration we accept 50 most confident predictions, but the number of accepted samples per activity needs to be matched to the initial distribution of activities in the training set. We performed experiments with fewer accepted samples per iteration, but in that case the learning phase is slower, because more iterations are required to achieve high performance. Additionally, in order to get more representative samples for the labeling process, as suggested in [3], we carried out random sampling of unlabeled training data and performed the labeling on that subset of data, but that did not improve the results.

Figures 2(a) and 2(b) show the mean and 95% confidence intervals of the classification accuracy of self-training (red bars) and co-training (green bars) when using different amounts of labeled training data for acceleration and infra-red sensors, respectively. The plots also show the comparison to the supervised approach (blue bars) when using the same decreased number of labeled training data, as well as the expected upper boundary (pink line) when using 100% of training data for supervised learning. From the plots one can clearly observe a superiority of co-training compared to self-training, e.g., when using 2.5% labeled training data

[1] These five configurations are constructed based on randomly sampled 50%, 25%, 10%, 5%, and finally only 1% of data from the selected two days. In each cross-validation round another two days of data are used for subsampling of labeled training set. As the amount of annotated data per day varies, these five configurations on average sums up to 12.5%, 6.3%, 2.5%, 1.3%, and 0.3% of the complete set of labeled and unlabeled training data.

Amount of labeled training data	100%	12.5%	6.3%	2.5%	1.3%	0.3%
Number of labels	9613	1203	604	244	124	29

Table 2. Average number of labels used for different experiment configurations

the performance of co-training is 12% higher than the performance of self-training on infra-red data. For acceleration data, accuracy increases by 9% when using the same amount of labeled training data, i.e., 2.5%. The performance of self-training on both sensor modalities, i.e., acceleration and infra-red does not differ significantly. For acceleration data there is a consistent improvement compared to the supervised approach with the same reduced amount of labeled training data. For infra-red data, after self-training the performance is sometimes degraded (when using 0.3% labeled training data), which highly depends on the quality of the initial labeled training subset of data. The experiments in Section 3 show that joint boosting performs better on acceleration data than on infra-red data. Therefore, the full strength of co-training is clear when looking at the benefit that infra-red data gain from co-training. The performance is boosted by more accurate acceleration predictions during co-training. In most of the configurations, it outperforms even the supervised approach when using 100% labeled training data. For the configuration when we use 2.5% labeled training data, the performance of co-training is 4% higher than in the supervised case of 100% labeled training data. Co-training of acceleration data never achieves the performance of the supervised case of 100% labeled training data, but the strength of the algorithm is still visible compared to the supervised case when using the same reduced amount of labeled training data as for co-training. In the case of 2.5% labeled training data, the increase of performance is 3% for self-training and 14.6% for co-training. Surprisingly, by using more labeled training data performance of co-training starts to decrease, presumably because of the noise in infra-red data that is more inherent in larger random subsets of data.

All the above mentioned results clearly show the potential of semi-supervised approaches to minimize the labeling efforts. As can be observed from Table 2, the number of labels averaged over 9 cross validation rounds is extremely reduced compared to the average of 9613 labels when using 100% labeled training data for supervised approach presented in Section 3. In the configuration when we use 2.5% labeled training data, as can be seen from Figures 2(a) and 2(b), the achieved results are impressive, considering that only 244 labels are used. In that case, 6 activity models are learned with less than 5 labels per activity. When further decreasing the number of labeled training samples, some of the activities are learned from a single label. In the extreme case, when using 0.3% labeled training data, i.e. only 29 labels, 6 out of 9 activities are learned from a single labeled sample per activity. In that case the achieved performance is relatively low due to the very few labels, but by carefully chosing the data to be labeled the performance can still be significantly improved. Therefore, in the next section we utilize active learning for activity recognition.

5. Active Learning Approach

Active learning aims at detecting the most informative unlabeled samples and queries a user to label them. In the context of activity recognition, one can legitimately imagine an online algorithm, similar to the stream-based setting in [8], that asks the user to annotate his current activity when it is considered necessary for improving the performance of recognition.

We employ a multi-sensor approach for active learning to select important samples to be labeled. The approach is based on a pool-based setting, i.e., we use a small set of labeled data and a large set of unlabeled data for training. The active learning algorithm searches for samples from the unlabeled training data to be labeled by a user. Two active sampling functions are evaluated here. The first function is based on the assumption that the most informative samples are those the classifiers are least confident about. The second function is based on the assumption that when the two classifiers have a high degree of disagreement about a certain sample, the sample should be labeled by a user.

More formally, let $h_c^1(x_i)$ and $h_c^2(x_i)$ be the two classifiers' confidence scores that sample x_i belongs to the class c based on two different sets of features. The first active sampling function asks for the label of the sample s_j with the lowest prediction score, i.e.,:

$$s_j = \underset{x_i}{\operatorname{argmin}}(\max_c h_c^j(x_i)),\ j = 1, 2 \tag{1}$$

The second active sampling function first finds the conflicts S in the classifiers' predictions:

$$S = \{x_i | \hat{c}_1(x_i) \neq \hat{c}_2(x_i)\} \tag{2}$$

where $\hat{c}_1(x_i)$ and $\hat{c}_2(x_i)$ are predicted classes:

$$\hat{c}_j(x_i) = \underset{c}{\operatorname{argmax}}\, h_c^j(x_i),\ j = 1, 2 \tag{3}$$

and then chooses for labeling the sample in the set S with the highest confidence score:

$$\underset{x_i \in S}{\operatorname{argmax}}(\max_j h_c^j(x_i)),\ j = 1, 2 \tag{4}$$

We evaluate the proposed active sampling functions based on the iterative training procedure. Again, we use 9-fold leave-one-day-out cross validation and 5 random subsampling rounds. We start with only a few labeled samples,

	Acceleration			Infra-red			Combined		
Labeled	Supervised	Active - low scores	Active - conflicts	Supervised	Active - low scores	Active - conflicts	Supervised	Active - low scores	Active - conflicts
1.3%	24.4% ± 5.7%	44.5% ± 1.6%	47.1% ± 3.0%	32.8% ± 7.5%	34.5% ± 3.8%	29.5% ± 3.2%	28.2% ± 3.4%	50.9% ± 1.8%	51.4% ± 3.4%
2.5%	26.9% ± 2.9%	52.9% ± 1.9%	51.4% ± 3.4%	30.5% ± 3.9%	39.8% ± 6.2%	38.0% ± 3.2%	32.3% ± 5.2%	59.7% ± 1.0%	57.0% ± 4.4%
6.3%	35.9% ± 4.3%	55.3% ± 2.4%	53.8% ± 3.1%	30.1% ± 3.8%	42.2% ± 1.8%	23.7% ± 4.4%	39.8% ± 4.3%	63.2% ± 1.6%	57.5% ± 2.0%
12.5%	38.9% ± 7.0%	**60.6% ± 2.3%**	55.8% ± 2.0%	24.5% ± 1.3%	**42.3% ± 2.1%**	32.2% ± 5.7%	35.8% ± 3.3%	**64.2 ± 1.9%**	63.5% ± 1.6%

Table 3. Comparison of recognition accuracy ± 95% confidence interval using 2 different active learning sampling functions and supervised learning for acceleration, infra-red, and combined classifier

i.e., with 0.3% labeled training data from the previous section. Joint boosting classifiers on acceleration and infra-red data are then trained and applied to the pool of unlabeled training data. The most informative samples are chosen for labeling by one of the two proposed active sampling functions and added to the labeled training set. The classifiers are then re-trained, and the procedure continues until the size of the labeled training data reaches the size of the four configurations from the previous section, i.e., 1.3%, 2.5%, 6.3%, and 12.5% of 8 days of training data.

In each iteration, the first active sampling function (Equation 1) finds two samples for labeling, the one that is predicted with the lowest confidence level based on the acceleration classifier, and the one that has the lowest score based on the infra-red classifier. These two samples are then labeled and added to the labeled training set. The second active sampling function (Equation 4) searches the prediction space for conflicts, i.e., samples that are classified differently by classifiers based on acceleration and infra-red data, and chooses for labeling the one that the classifiers predicted with the highest confidence level. That sample is then labeled and added to the set of labeled training data.

5.1. Results

Table 3 shows the classification results for acceleration and infra-red data, as well as for the classifier combined on these two sensor modalities, after the active sampling labeling process. We compare the results for different amounts of data sampled with the two previously introduced active sampling functions. Additionally, the results are compared with the supervised approach when using the same amount of non-actively (i.e. randomly) sampled labeled training data. Both active sampling functions outperform the supervised learning approach. On average, the first active sampling function for acceleration data based on the low confidence predictions' scores yields 20.6% better accuracy, and the second active sampling function based on conflicts in classifiers' predictions achieves 21.5% better acccuracy compared to the supervised case with the same amount of labeled training data. In the case of infra-red data the performance increase is less significant, but still noticeable. Again, we assume that this is due to the noise in the infra-red data introduced by the second subject, which joint boosting can not deal with properly. The active sampling function based on the low predictions' scores after labeling 6.3% and 12.5% of training data achieves an accuracy of 42.2% and 42.3%, respectively, which is slightly better compared even to the supervised learning by using 100% of labeled infra-red training data when accuracy is 41.6%.

One must be aware of the potential risk that active learning might focus on the samples that are hard to be learned. It happens occasionally that accuracy decreases by adding more actively sampled labels. For example, when using the active sampling function based on the conflicts for infra-red data accuracy is 38% when 2.5% data is labeled. By continuing the active labeling and reaching 6.3% labeled data, accuracy decreases to 23.7%.

In order to explore the full potential of the multi-sensor approach, in Table 3 we also show the performance of the combined classifier, based on the multiplied outputs from the acceleration and infra-red classifiers. That way, we achieve an accuracy of 64.2% when the active sampling function based on the low prediction scores is used and 63.5% when using the active sampling function based on the classifiers' prediction conflicts. In Table 3, the best results for acceleration, infra-red and combined classifier are highlighted and the active sampling function based on the low prediction scores consistently performs better, presumably because the active sampling function based on conflicts in classifier's prediction often chooses for labeling the samples close to the decision boundaries.

When comparing the three approaches used in this paper, one can conclude that the most promising approach is the combined classifier on the actively learned data. Table 4 ranks the best results for sensor modalities separately.

6. Conclusions and Future Work

This paper demonstrated the feasibility of semi-supervised and active learning for reducing the level of supervision in activity recognition.

The two evaluated semi-supervised techniques, self-training and co-training, were found to be capable of learn-

Acceleration		Infra-red	
Active - low scores	60.6%	Co-training	45.9%
Active - conflicts	55.8%	Active - low scores	42.3%
Supervised	53.6%	Supervised	41.6%
Co-training	40.7%	Active - conflicts	38.0%
Self-training	40.6%	Self-training	34.4%

Table 4. Comparison of the best recognition accuracy for all the approaches used

ing activity models from a very limited amount of labeled training data. As intuitively assumed, experimental results showed that co-training outperforms self-training by augmenting the training process with additional information from complementary sensor modalities. Additionally, in some cases it can achieve higher recognition accuracy than the fully supervised approaches.

The proposed active learning method is based on a pool-based setting where in addition to a small set of labeled training data, there is also a large number of unlabeled training instances available. From the unlabeled pool of data, the algorithm selects the most informative samples to be labeled by user. We introduced two active sampling functions based on the classifiers' lowest confidence level and on disagreements between the classifiers' predictions. Again, experimental results suggest that it is possible to achieve comparable, or sometimes even higher accuracy than the fully supervised approaches with less labeling efforts.

In the future, we plan to investigate a hybrid approach that would in the initial phase actively ask for labels of the most profitable samples. In the second phase, co-training could highly benefit from actively learned labels.

7. Acknowledgments

We would like to thank Beth Logan, Jennifer Healey, Emmanuel Munguia Tapia, and Stephen Intille for their assistance and providing access to the PLCouple1 dataset. This research has been supported in part by Fraunhofer's Doktorandinnenprogram.

References

[1] B. Anderson and A. Moore. Active Learning for Hidden Markov Models: Objective Functions and Algorithms. In *ICML'05*.

[2] L. Bao and S. Intille. Activity Recognition from User-Annotated Acceleration Data. In *Pervasive'04*.

[3] A. Blum and T. Mitchell. Combining Labeled and Unlabeled Data with Co-Training. In *COLT'98*.

[4] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[5] D. Guan, W. Yuan, Y.-K. Lee, A. Gavrilov, and S. Lee. Activity Recognition Based on Semi-supervised Learning. In *RTCSA'07*.

[6] T. Huynh and B. Schiele. Unsupervised Discovery of Structure in Activity Data using Multiple Eigenspaces. In *LoCA'06*.

[7] S. S. Intille et al. Using a Live-In Laboratory for Ubiquitous Computing Research. In *Pervasive'06*.

[8] A. Kapoor and E. Horvitz. Experience Sampling for Building Predictive User Models: A Comparative Study. In *CHI'08*.

[9] A. Krause, D. Siewiorek, A. Smailagic, and J. Farringdon. Unsupervised, Dynamic Identification of Physiological and Activity Context in Wearable Computing. In *ISWC'03*.

[10] K. V. Laerhoven, N. Kern, H.-W. Gellersen, and B. Schiele. Towards a Wearable Inertial Sensor Network. In *IEE Eurowearable'03*.

[11] J. Lester, T. Choudhury, and G. Borriello. A Practical Approach to Recognizing Physical Activities. In *Pervasive '06*.

[12] B. Logan, J. Healey, M. Philipose, E. Tapia, and S. Intille. A Long-Term Evaluation of Sensing Modalities for Activity Recognition. In *Ubicomp '07*. Dataset http://architecture.mit.edu/house_n/data/PlaceLab/PlaceLab.htm.

[13] M. Mahdaviani and T. Choudhury. Fast and Scalable Training of Semi-Supervised CRFs with Application to Activity Recognition. In *NIPS'07*.

[14] D. Minnen, T. Starner, I. Essa, and C. Isbell. Discovering Characteristic Actions from On-Body Sensor Data. In *ISWC'06*.

[15] I. Muslea, S. Minton, and C. A. Knoblock. Selective Sampling with Redundant Views. In *AAAI/IAAI*, 2000.

[16] N. Ravi, N. Dandekar, P. Mysore, and M. Littman. Activity Recognition from Accelerometer Data. In *IAAI'05*.

[17] T. Stiefmeier, G. Ogris, H. Junker, P. Lukowicz, and G. Tröster. Combining Motion Sensors and Ultrasonic Hands Tracking for Continuous Activity Recognition in a Maintenance Scenario. In *ISWC'06*.

[18] A. Subramanya, A. Raj, J. Blimes, and D. Fox. Recognizing Activities and Spatial Context Using Wearable Sensors. In *UAI'06*.

[19] E. M. Tapia, S. S. Intille, and K. Larson. Activity Recognition in the Home Using Simple and Ubiquitous Sensors. In *Pervasive'04*.

[20] E. M. Tapia, S. S. Intille, L. Lopez, and K. Larson. The Design of a Portable Kit of Wireless Sensors for Naturalistic Data Collection. In *Pervasive'06*.

[21] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *CVPR'04*.

[22] S. Wang, W. Pentney, A.-M. Popescu, T. Choudhury, and M. Philipose. Common Sense Based Joint Training of Human Activity Recognizers. In *IJCAI'07*.

[23] J. Ward, P. Lukowicz, and G. Tröster. Gesture Spotting Using Wrist Worn Microphone and 3-Axis Accelerometer. In *sOc-EUSAI*, 2005.

[24] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[25] D. Wyatt, M. Philipose, and T. Choudhury. Unsupervised Activity Recognition Using Automatically Mined Common Sense. In *AAAI'05*.