From on-going to complete activity recognition exploiting related activities

Carlo Nicolini[1], Bruno Lepri[2], Stefano Teso[1], and Andrea Passerini[1]

[1] Dipartimento di Ingegneria e Scienza dell'Informazione
Università degli Studi di Trento, Italy
[2] FBK-irst, via Sommarive 18, Povo, Trento, Italy

Abstract. Activity recognition can be seen as a local task aimed at identifying an *on-going* activity performed at a certain time, or a global one identifying time segments in which a certain activity is being performed. We combine these tasks by a hierarchical approach which locally predicts on-going activities by a Support Vector Machine and globally refines them by a Conditional Random Field focused on time segments involving related activities. By varying temporal scales in order to account for widely different activity durations, we achieve substantial improvements in on-going activity recognition on a realistic dataset from the PlaceLab sensing environment. When focusing on periods within which related activities are known to be performed, the refinement stage manages to exploit these relationships in order to correct inaccurate local predictions.

1 Introduction

Automatic monitoring of Activities of Daily Living (ADLs, such as eating, drinking, cleaning, and so on) is an important component for the implementation of advanced services in the fields of Ambient Assisted Living and Assisted Cognition. In assessing the level of self-sufficiency of patients, clinicians consider the capabilities of performing basic ADLs such as cooking and eating [1]. The automatic recognition and tracking of these activities may allow for a more reliable and cheaper automatic reporting to clinicians or relatives. At the same time, it allows for the provision of advanced services that can contribute to older people's independent life: services like reminders, help in activity execution, etc.

As defined in [2], the activity classification task can take at least two guises which differ according to the kind of perspective taken on the activities. The first type is the "complete activity" (CA) recognition task and considers finished activities and asks about their type. This task involves an external perspective on the activity and humans talk about these activities using the perfective tenses as in the following example: A-"What did Mark do yesterday afternoon?", B-"He played basketball". For automatic systems, the task is to assign the right activity label to the unknown segmented and complete one. Different works in activity recognition field dealt with CA task; in particular it has been often used by researchers adopting the object-use approach whereby activities are modeled as sequences of used objects [3,4]. The second kind of classification task, called "on-going activity" recognition task (OGA) takes an internal perspective on activities. The human subject or the automatic system are temporally located inside the activity. In this case, humans would use imperfective tenses or progressive forms: A-"What is Mark doing now?", B-"Mark is playing basketball". So, the task is anchored to a given time and the goal of the human or of the machine is finding signs of the on-going activity and define their type. We can cite some previous works adopting the OGA paradigm; for example, [5,6]. In this paper we are going to deal with both of these tasks (OGA and CA): more precisely, we deal with OGA task using Support Vector Machines (SVM) in order to predict what is happening inside a given small time interval. These local predictions are fed as input to a sequential model, namely a Conditional Random Field (CRF), aimed at performing CA recognition on larger segments of the day. In the real world, people often perform multiple activities concurrently in their daily living; e.g. a person might have the habit of watching TV while ironing. Furthermore, related activities can have quite different recognition complexity. We build on this observation by focusing on time segments involving highly related activities, and exploiting a well-predicted activity to improve recognition of a difficult one. Finally, our evaluation highlights the importance of calibrating the temporal scale at which an activity should be searched for depending on its average duration.

This paper is organized as follows. In Section 2 we discuss some related works on activity recognition. Section 3 describes the sensing environment and the learning algorithms we employed. Experimental results are reported in Section 4, and conclusions are drawn in Section 5.

2 Previous works

The problem of human activity recognition has received increasing interest in recent years in the pattern recognition and machine learning communities. In particular, good results were achieved both on low-level activities (e.g. ADLs such as sitting, standing, walking, and lying [7,8]), and high-level activities (e.g. eating, watching TV, dishwashing, and cooking [9,10], and office activities [11]). Different sensors were used for activity recognition tasks: several works have explored the use of switches and motion detectors (similar to those used in common alarm systems) to collect data regarding the performance of ADLs [12]. Recently, Logan et al. [5] compared different modalities on data approaching real-world conditions: they collected 104 hours of annotated data of a person living in a house, instrumented with over 900 sensors, including power and water flow inputs, objects and person motion detectors, and RFID tags. They found that 10 infra-red motion detectors outperformed the other sensors on many of the studied activities, especially those that were usually performed in the same location. From a machine learning point of view, most of the work in the activity recognition area is based on supervised algorithms such as Naive Bayes [9], Decision Trees [5,7], Hidden Markov Models [3,8,13,14], Support Vector Machines [2, 13], and Conditional Random Fields [14]. In particular, Conditional Random Fields were found to offer higher overall accuracy than Hidden Markov Models (HMM) for multi-label activity classification, even if HMMs can better discriminate between multiple activities when the training dataset contains unbalanced class labels [14]. A limited number of works used relational learning techniques to deal with activity recognition tasks: [15] used Relational Markov Networks (RMNs) for recognizing activities from location data. Landwher et al [16] introduced a relational transformation based tagging system in order to integrate various principles of inductive logic programming (e.g., search, operators, representations, and background knowledge) with transformation-based tagging (e.g., error-driven search, branch and bound idea).

3 Activity Classification

3.1 The sensing environment

PlaceLab is an instrumented home environment operated as a shared research facility. The complete description of the sensing environment can be found in [5]. Logan et al. [5] collected and analyzed data from a couple who lived at the home for a period of 10 weeks. The home is a custom built condominium instrumented with several hundred sensors, including an audiovisual recording system that captures ground truth of the participants activities. The environment contains several classes of sensors, including wired reed switches, power and water flow inputs, objects and person motion detectors, and RFID tags. We focused on infrared (IR) and object motion (OM) sensors, those found in Logan et al. [5] to be the most discriminant.

Table 1. Activity duration statistics

Activity	Instances	Avg. duration (min)
ActivelyWatchingTV	15	53
DishWashing	21	1
Grooming	28	2
GroupedEating	101	4
Hygiene	20	3
MealPrep	40	2
Reading	29	17
UsingComputer	50	37
UsingPhone	68	3

3.2 Data Preparation

Following Logan et al. [5], we divided each day into $30s$ intervals overlapped by $15s$. We formulated the activity recognition problem as nine binary classification tasks at the interval level, one for each of the nine possible activities.

Each interval was labeled positively for a certain activity if it had occurred at any time within it. Note that many activities are not mutually exclusive (i.e. they can both be performed within the $30s$ timeframe) and the problem should be addressed as multi-label rather than multi-class prediction.

We represented an interval as a vector of sensor features, indicating the number of times each sensor was activated during the interval.

Most activities have average durations on the order of a minute, with some like ActivelyWatchingTV or UsingComputer having a far longer duration as showed in Table 1.

In order to account for such temporal correlations we applied a sliding window approach computing average feature vectors on intervals surrounding the one of interest, either separately for past and future (asymmetric) or combining both together (symmetric).

3.3 Local classification by Support Vector Machines

We addressed each binary classification task at the interval level with an SVM classifier [17]. SVM are state-of-the-art discriminative classifiers capable of efficiently handling thousands of features and learning complex non-linear functions thanks to the kernel trick. Experimental results show substantial improvements over the decision tree classifiers employed in [5], as will be detailed in the experimental section.

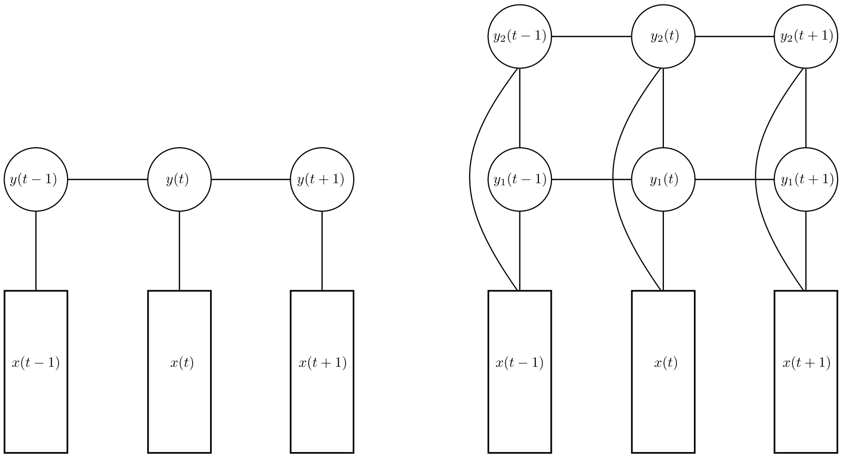

Fig. 1. Graphical model representation of a linear-chain CRF, on the left, and a factorial CRF with two chains, on the right, unrolled for three time intervals. The $x(t)$ nodes represent the predicted OGAs over time, while the $y_i(t)$ variables are the detected CAs.

3.4 Global refinement by Conditional Random Fields

CRFs [18] are undirected graphical models conditioned on observation sequences. Linear-chain CRF allow to efficiently model sequential observations and have been successfully applied to a variety of recognition tasks in text classification, bioinformatics and activity recognition, to name a few application domain. Here we employ them as a *refinement* stage, in order to combine sequences of local OGA predictions from multiple related activities into a global CA prediction. Figure 1 shows a graphical representation of the models we employed. The inputs $x(t)$ represent local OGA predictions for all or some of the activities at time interval t. The outputs $y(t)$ represent CA predictions for the activity being globally refined. The model to the left is a plain linear-chain CRF, where a single activity is predicted in output. Connections are provided between outputs at consecutive time instants, with the effect of propagating predictions along the time range. The model to the right is a more complex factorial CRF [19], where multiple activities (two in this example) are jointly predicted. Linear-chain models for each activity are combined by adding co-temporal connections between activities. Note that the higher complexity of the factorial model implies more parameters to be estimated and approximate inference. In conjunction with the scarcity of positive examples for most activities, this often resulted in a performance worsening with respect to the simpler linear-chain case, as will be detailed in the experimental section.

4 Experimental results

We conducted a leave-one-day-out cross validation procedure as in [5]. The aims of the experimental evaluation are: 1) identifying the most discriminative sensors and time frames (i.e. sizes of the sliding windows) for the different activities; 2) comparing to previous activity recognition approaches on this dataset; 3) verifying the usefulness of sequential models to refine local predictions. In the following we will report experimental results for each of these points. For comparability to [5], we employed area under the ROC curve (AUC) as a figure of merit in all experiments.

For each of the binary classification tasks, we conducted an extensive model selection phase to identify 1) the best set of sensors, IR, OM or IR+OM; 2) the best sliding window size; 3) the best SVM parameters, namely regularization parameter C and kernel type among linear, polynomial or Gaussian with varying width size.

Model selection was conducted by an inner leave-one-day-out cross validation on the training set of the first fold (i.e. the first 8 days), and obtained parameters and feature sets were kept fixed for the outer cross validation.

4.1 Model selection results

The best kernel was a second degree polynomial for all activities. Concerning sensor classes, IR sensors performed much better than OM ones. Furthermore,

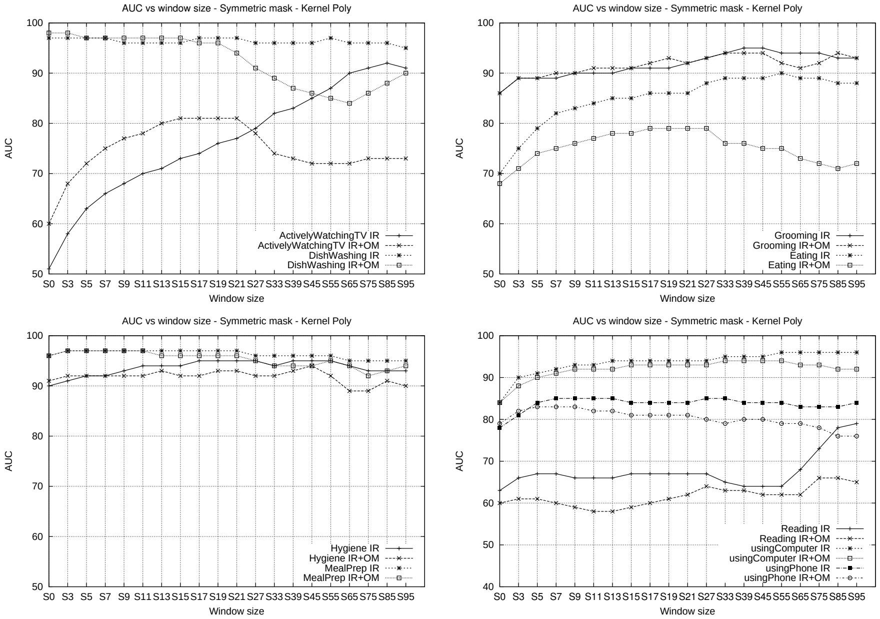

Fig. 2. AUC dependence on window size, comparison between IR and IR+OM for the different activities.

we did not experience significant advantages in combining OM and IR sensors, especially when increasing the size of the sliding window. These results are consistent with those reported in [5] where IR sensors where found to be the most discriminant.

Figure 2 reports AUC values for varying window sizes for the different activities. Both IR and IR+OM results are shown. Two aspects are worth mentioning. First, large differences can be observed in the optimal window size of different activities. This size is actually highly correlated with the average duration of the activity (see Table 1), with the three longest activities, namely ActivelyWatchingTV, Reading (for which the maximum is outsize of the range shown) and UsingComputer, having by far the largest optimal window sizes. Second, IR and IR+OM behave quite differently with respect to optimal window size, with the latter early starting to show performance worsening. This seems to indicate the need to separately optimize window sizes for the two classes of sensors. We plan to investigate this issue in future experiments.

4.2 SVM results

Table 2 reports experimental comparisons between our local SVM classifiers and the decision trees (DT) used by Logan et al. [5].

SVM substantially outperforms DT in all experiments. The largest improvements can be observed for the three hardest recognition tasks, GroupedEating,

Table 2. Leave-one-day-out cross validated AUC (%) for DT and SVM. Optimal window size refers to SVM and was obtained by an inner cross-validation on the training set of the first fold.

Activity	Best win.	SVM$_{AUC}$	DT$_{AUC}$
ActivelyWatchingTV	$s85$	**90**	80
DishWashing	$s3$	**97**	89
Grooming	$s45$	**95**	87
GroupedEating	$s55$	**91**	56
Hygiene	$a19$	**96**	86
MealPreparation	$s11$	**97**	87
Reading	$s135$	**81**	54
UsingComputer	$s95$	**96**	85
UsingPhone	$s9$	**85**	64

Reading and UsingPhone. Note that an appropriate window size is also crucial in achieving these results, especially for the first two activities which perform drastically worse if only the activations in the target interval are considered (see Figure 2). UsingComputer is by far the best predicted activity, as the other activities with AUC > 0.95 have much less positive examples, and AUC is very sensitive to the unbalancing in the data.

4.3 CRF results

We investigated the usefulness of relying on well-predicted activities in order to improve recognition of more difficult ones. Figure 3 shows the cross-correlation between UsingComputer and Reading, the two activities with highest cross-correlation. Note that activities are frequently co-occurrent (Lag=0), but also frequently follow one after the other within a short time frame.

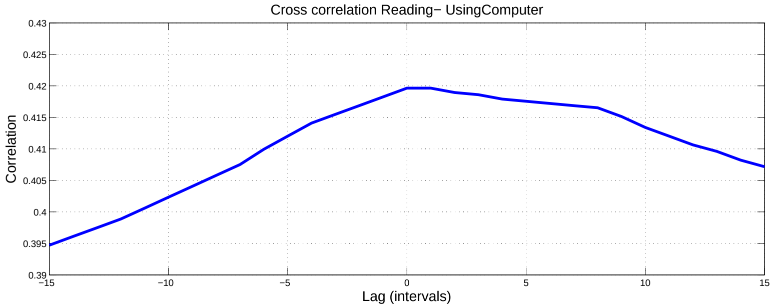

Fig. 3. Cross correlation between Reading and UsingComputer labels

We employed either the true labels or the local SVM predictions in order to focus on time segments likely to contain one of these activities. We selected segments of consecutive intervals where at least one of the two activities was actually performed or locally predicted to be performed, allowing for a small gap of inactivity (10 intervals) between consecutive positive intervals. We then retained those segments in which each activity was performed for at least 3 intervals. During test, we applied the same selection mechanism to identify candidate time segments.

We experimented with two different models: a linear-chain CRF predicting a single activity, and a factorial CRF jointly predicting Reading and UsingComputer. Each model was input either the margins of the activity being predicted, the margins of both activities, or the margins of all the nine activities.

Table 3. Results of CRF experiments. Both training and test data is segmented according to the true labels. A '-' indicates that there are no positive instances of the given activity. L-CRF stands for linear-chain CRF, F-CRF for factorial CRF.

Predictions for Reading									
Prediction	1	2	3	4	5	6	7	8	9
SVM	0.31	–	0.25	0.48	**0.85**	–	0.12	0.55	0.56
L-CRF, Reading	0.38	–	0.5	0.85	0	–	0.66	0.32	0.56
L-CRF, Reading+UsingComputer	0.38	–	0.5	**0.94**	0	–	**0.86**	0.5	0.54
L-CRF, All Activities	**0.45**	–	0.5	0.71	0	–	0.67	**0.68**	0.65
F-CRF, Reading+UsingComputer	0.36	–	0.04	0.61	0.24	–	0.33	0.49	0.48
F-CRF, All Activities	0.42	–	**0.93**	0.55	0	–	0.69	0.5	**0.72**
Predictions for Using Computer									
Prediction	1	2	3	4	5	6	7	8	9
SVM	0.48	–	0.95	0.68	0.8	–	0.79	0.83	0.77
L-CRF, UsingComputer	0.99	–	0.98	0.94	0.9	–	**0.91**	0.76	**0.89**
L-CRF, Reading+UsingComputer	0.98	–	0.98	0.94	0.9	–	0.87	0.78	**0.89**
L-CRF, All Activities	**1**	–	**1**	0.85	**0.97**	–	**0.91**	**0.87**	0.87
F-CRF, Reading+UsingComputer	0.71	–	0.95	0.86	0.78	–	0.79	0.86	0.79
F-CRF, All Activities	0.72	–	0.67	**0.96**	0.14	–	0.87	**0.87**	0.78

Table 3 summarizes the results for the prediction of Reading and UsingComputer with the different CRF models. Each row contains the AUC of the predictions for the given combination of CRF type and inputs, for each day of the test data. The AUCs of the SVM predictor are included for reference. Numbers in bold highlight the best result for each test day. Using the true labels to segment both train and test data is clearly infeasible in realistic conditions, where the true labels are not available during the testing stage. However, these experiments allow us to highlight the potential advantages of a sequential refinement stage for CA recognition, abstracting away the problem of identifying candidate periods of the day to focus on.

The first observation is that the linear-chain CRF typically outperforms the SVM, on all days and for both activities, with the sole exception of Reading during day 5. In general the CRF manages to overcome the local predictions.

These can be very bad especially in the case of Reading, which is particularly difficult to predict on a local basis. Interestingly, this behavior occurs even if only one input is given. A particular instance of this behavior can be seen in Figure 4, referring to day 7 of the one input case. Here the local predictions are quite bad, but the CRF is able to approximately detect the second large segment of activity.

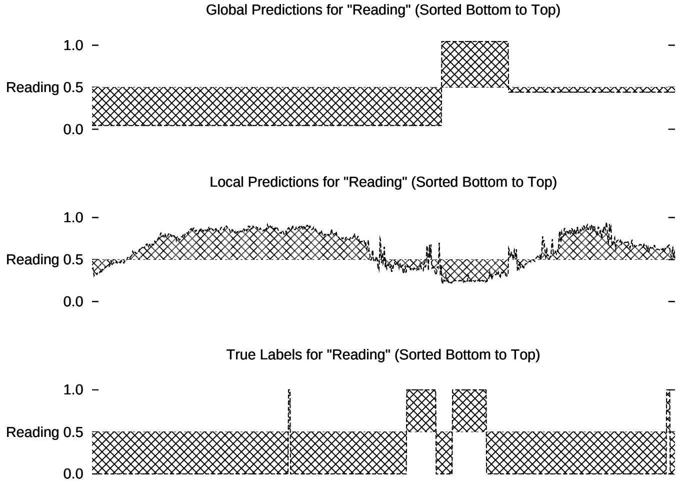

Fig. 4. Plot showing the predictions of the linear-chain CRF for Reading on cross-validation day 7. On the bottom row we report the true labels for Reading. The middle row represents the SVM predictions. The top represents the CRF predictions.

The factorial CRF does not show this consistent behavior, performing rather worse on some of the test instances. This may be due to the higher complexity of the model, requiring more parameters and approximate inference, and the sparseness of the train data available for Reading. One exception is shown in Figure 5, which refers to the predictions from all activities as inputs on day 4.

We also note that usually increasing the number of inputs improves the prediction for both the linear-chain and factorial models. This fact hints at the positive effect that combining multiple local predictions has on the accuracy of the CRF. As an example, Figure 6 shows how the CRF combines the wrong local prediction of Reading with the prediction of UsingComputer to accurately locate both positives and negatives of Reading, even though its SVM prediction is almost completely wrong.

Table 4 summarizes the results of the CRFs when the test and train data are segmented according to the local predictions. In this case, the contributions of the CRF are not as clear cut. For both activities, the best CRF model seems the most complex one, namely a factorial CRF with all 9 activities in input. However,

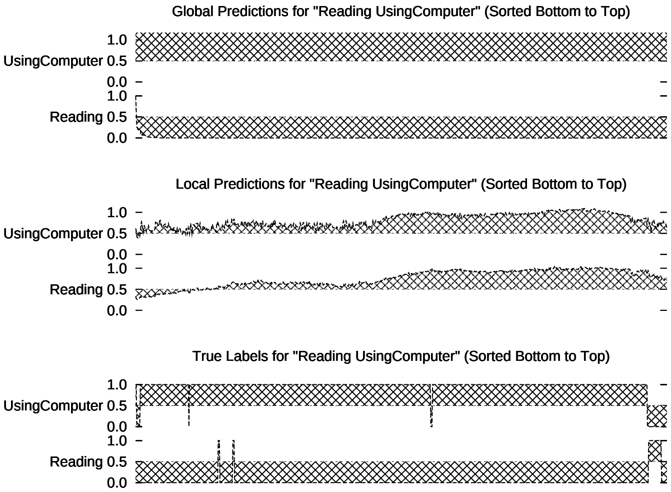

Fig. 5. Plot showing the predictions of the factorial CRF for both Reading and UsingComputer, with all 9 activities as inputs. The predictions refer to day 4. For simplicity, only the labels and local predictions for Reading and UsingComputer are shown.

the comparison with the local SVM predictions does not allow to draw clear conclusions, with three wins vs three losses for3.4 Reading, and five wins vs three losses for UsingComputer. Experiments in which training data were segmented according to the true labels did not produce substantially different results. This indicates that further work is needed in order to make CRF predictions more robust to a noisy identification of candidate periods.

5 Conclusion

We addressed the problem of activity recognition from the two perspectives of on-going and complete identification. We showed that by varying the temporal scale at which sensor readings are aggregated, we can account for the different average duration of activities, achieving substantial improvements on the on-going recognition task. The combination of local predictions by CRF sequential models allowed us to refine them into a complete activity recognition prediction. Preliminary results indicate that when focusing on periods containing related activities, this relationship helps to correct inaccurate local predictions, especially in exploiting information on easier activities to improve predictions of a harder one. In order to successfully apply this strategy in a real setting, however, we need to improve its robustness to a noisy identification of these periods, for instance by focusing on reliable predictions only and searching for the more difficult activities in the surroundings of the simpler ones.

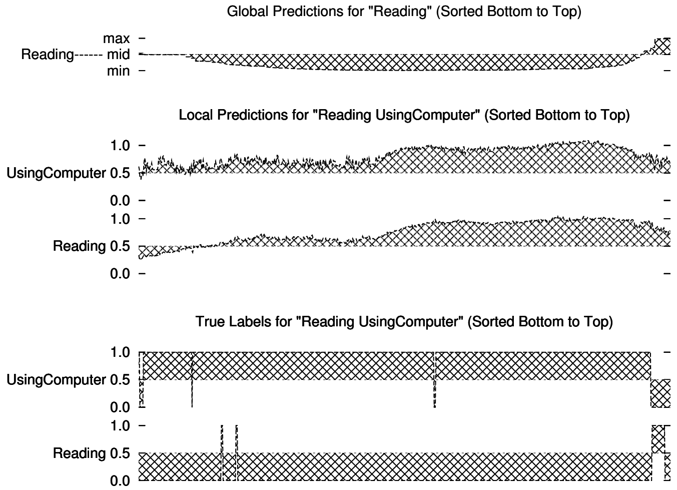

Fig. 6. Plot showing the predictions of Reading with both Reading and UsingComputer as inputs, performed with the linear-chain CRF. The predictions refer to day 4.

Aknowledgments

The research was funded by the Autonomous Province of Trento, Call for proposal Major Projects 2006 (project ACube).

References

1. Katz, S.: Assessing self-maintenance: Activities of daily living, mobility, and instrumental activities of daily living. Journal of American Geriatrics Society **31**(12) (1983) 712–726
2. Lepri, B., Mana, N., Cappelletti, A., Pianesi, F., Zancanaro, M.: What is happening now? detection of activities of daily living from simple visual features. Personal and Ubiquitous Computing (2010)
3. Philipose, M., P, K., Perkowitz, M., Patterson, D.J., Fox, D., Kautz, H., Hähnel, D.: Inferring activities from interactions with objects. IEEE Pervasive Computing **3** (2004) 50–57
4. Pentney, W., Philipose, M., Bilmes, J.A., Kautz, H.A.: Learning large scale common sense models of everyday life. In: AAAI. (2007) 465–470
5. Logan, B., Healey, J., Philipose, M., Tapia, E.M., Intille, S.: A long-term evaluation of sensing modalities for activity recognition. Proceedings of the International Conference on Ubiquitous Computing **LNCS 4717** (2007) 483–500
6. Stikic, M., Huynh, T., Van Laerhoven, K., Schiele, B.: Adl recognition based on the combination of rfid and accelerometer sensing. In: 2nd International Conference on Pervasive Computing Technologies for Healthcare 2008. (2008)
7. Bao, L., Intille, S.S.: Activity recognition from user-annotated acceleration data. In: Pervasive 2004, Springer (2004) 1–17
8. Lester, J., Choudhury, T., Borriello, G.: A practical approach to recognizing physical activities. In: Proc. of Pervasive. (2006) 1–16

Table 4. Results of CRF experiments. Both training and test data is segmented according to the local predictions. A '-' indicates that there are no positive instances of the given activity.

Predictions for Reading									
Prediction	1	2	3	4	5	6	7	8	9
SVM	–	–	–	**0.62**	**0.95**	**0.76**	0.33	0.91	0.74
L-CRF, Reading	–	–	–	0.49	0.42	0.35	0.78	0.1	0.49
L-CRF, Reading+UsingComputer	–	–	–	0.51	0.39	0.23	**0.79**	0.05	0.39
L-CRF, All Activities	–	–	–	0.43	0.36	0.47	0.71	0.63	**0.86**
F-CRF, Reading+UsingComputer	–	–	–	0.46	0.08	0.33	0.71	0.05	0.36
F-CRF, All Activities	–	–	–	0.14	0.02	0.45	0.4	**0.94**	0.8
Predictions for Using Computer									
Prediction	1	2	3	4	5	6	7	8	9
SVM	**0.88**	0.79	–	0.83	**0.96**	0.39	0.88	0.79	0.55
L-CRF, UsingComputer	0.6	**0.97**	–	0.92	0.88	0.48	0.86	**0.9**	0.21
L-CRF, Reading+UsingComputer	0.6	0.88	–	0.85	0.76	0.55	0.83	**0.9**	0.14
L-CRF, All Activities	0	0	–	0.4	0.4	0.47	0.65	0.65	**0.86**
F-CRF, Reading+UsingComputer	0.38	0.85	–	0.67	0.81	**0.67**	0.21	0.48	0.43
F-CRF, All Activities	0.69	0.53	–	**0.99**	0.93	0.48	**0.94**	0.88	0.64

9. Tapia, E.M., Intille, S.S., Larson, K.: Activity recognition in the home using simple and ubiquitous sensors. In: Pervasive'04. (2004) 158–175
10. Wyatt, D., Philipose, M., Choudhury, T.: Unsupervised activity recognition using automatically mined common sense. In: AAAI. (2005) 21–27
11. Oliver, N., Horvitz, E., Garg, A.: Layered representations for human activity recognition. In: Fourth IEEE Int. Conf. on Multimodal Interfaces. (2002) 3–8
12. Ogawa, M., Ochiai, S., Shoji, K., Nishihara, M., Togawa, T.: An attempt of monitoring daily activities at home. In: 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society. (2000) 786–788 vol.1
13. Blanke, U., Schiele, B.: Scalable recognition of daily activities with wearable sensors. In: 3rd International Symposium on Location- and Context-Awareness (LoCA), Oberpfaffenhofen, Germany, Springer (September 2007)
14. van Kasteren, T., Noulas, A., Englebienne, G., Kröse, B.: Accurate activity recognition in a home setting. In: UbiComp '08, New York, NY, USA, ACM (2008) 1–9
15. Liao, L., Fox, D., Kautz, H.: Location-based activity recognition using relational markov networks. In: IJCAI'05. (2005)
16. Landwher, N., Gutmann, B., Thon, I., Philipose, M., De Raedt, L.: Relational transformation-based tagging for human activity recognition. In: Proceedings of the 6th Workshop on Multi-Relational Data Mining (MRDM). Warsaw, Poland. (September 2007)
17. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines. Cambridge University Press (2000)
18. Sutton, C., Mccallum, A.: Introduction to conditional random fields for relational learning. In Getoor, L., Taskar, B., eds.: Introduction to Statistical Relational Learning. MIT Press (2006)
19. yu Wu, T., chun Lian, C., jen Hsu, J.Y.: Joint recognition of multiple concurrent activities using factorial conditional random fields. In: 2007 AAAI Workshop on Plan, Activity, and Intent Recognition. (2007)