Location based Hierarchical Approach for Activity Recognition with Multi-modal Sensor Data

Xi Liu[1], Lei Liu[2], and Pang-Ning Tan[1]

[1] *Department of Computer Science and Engineering, Michigan State University*
[2] *Hewlett-Packard Laboratories*

{liuxi4,ptan}@cse.msu.edu, lei.liu2@hp.com,

Abstract. Accurate identification of human activities could help us provide a better patient recovery training guidance, or an early alarm of emergency that may happen to elder people, such as stroke, falls, etc. In this paper, we describe a hierarchical structure for indoor human activity recognition by team **MSUHPDM** during ECML/PKDD 2016 Sphere Challenge. The challenge requests a method to predict human activity through data fusion and pattern recognition from a common platform of non-medical sensors in the home environment. Instead of training a prediction model that learns decision surface for all activity classes together, we build a location-based hierarchy to organize such activity classes for prediction purpose. The first layer will partition the activity labels into different branches, one for each location. The second layer contains all the leaf nodes, where each node associates with classes that happen in a specific location. Local classifiers are then trained and ensembled to get the final prediction. Additionally, temporal activity consistency is also involved for enhancement. Our experiment shows that the proposed approach is promising and effective.

Keywords: hierarchy; sensor; human activity

1 Introduction

In this paper, we aim to recognize 20 human activities from wrist-worn accelerometer, RGB-D camera and Passive Infrared (PIR) sensor data, all of which are generated from mature sensing technologies, that are broadly built in off-the-shelf smart devices. The human activities that we considered can be generally cast into three groups: **Motions**: activities requiring continuing movement, including *ascend stairs, descend stairs, jump, walk with load* and *walk*; **Stationary postures**: times when the participants are stationary, including *bending, kneeling, lying, sitting, squatting* and *standing*; **Transitions**: describe posture-to-posture transition activities, including *stand-to-bend, kneel-to-stand, lie-to-sit, sit-to-lie, sit-to-stand, stand-to-kneel, stand-to-sit, bend-to-stand* and *turn*.

Given wrist-worn accelerometer, RGB-D camera and PIR data that is associated with each single second, our goal is to predict the probability of each

human activity within each target second. To summarize, we have the several challenges in this task. **First**, missing values appear in both accelerometer and RGB-D camera data. The strategy of directly discarding seconds that contain missing values will leave us too few examples to train an effective predictive model. How to effectively handle such data missing issue is the first problem we need to solve. **Second**, a single second may be associated with more than one activity. Therefore, instead of assigning only a single activity to each second, we need to learn a model, which has the capability of assigning probability scores to all possible activities within each target second. **Third**, human activities happen in continuous time, such temporal relations between different activities is an crucial information that we should not discard. For example, *lying* rarely happens next to *jumping*; *sitting* rarely happens next to *ascending stairs*; and *lie-to-sit* usually happens within *lie* and *sit*.How to effectively incorporate such temporal relationships in our algorithm is also a challenge. **Furthermore**, location information is important to our task, and how to effectively import such prior knowledge of activities at different locations is crucial. For example, it is more likely that the subject is ascending or descending the stairs rather than other activities on the stairs. **Finally**, there exits serious class unbalance problem in such sensory data, because in a majority of time, the observations of activities belong to the a few largest classes such as *sitting, standing* than small classes, such as *squat, bend*. Due to this class unbalance problem, standard classifier tends to be more biased towards assigning observations into larger classes rather the smaller classes, which however are usually more of our interest. How to deal with such class imbalanced data is another challenge that must be addressed.

To predict the activity label for each target second, instead of learning a global prediction model for all activity classes, we learn a hierarchical model by leveraging the location-activity relational information. Our method has several advantages. First, it benefits the prediction performance by leveraging the relational information between location and each activity class, such as ascend/descend can only happen in hallway and not other places, in our prediction framework. Second, Our hierarchical approach could alleviate the class imbalanced problem by grouping together the related activity classes. Third, we train a local classifier at each leaf node, which associate less number of classes, this will allow us to have a less training complexity when compared against training over all the classes in a joint model. Fourth, our method considers activity temporal information by leveraging the activity information of connecting seconds of the target one. Furthermore, our model has the capability of assigning probability scores to all possible activities within each target second.

2 Preliminaries

2.1 Overview of the data

In this paper, we mainly investigated the labeled data from "SPHERE Challenge: Activity Recognition with Multimodal Sensor Data" [1] for indoor human activity recognition (denoted as "Sphere data" in the rest of the text). The Sphere

data provides us 10 sequences of labeled sensory data with manual annotations. The sensors used include:

- An accelerometer on the wrist of a subject
- An RGB-D camera placed each in hallway, kitchen, and living room
- A passive environmental sensor (PIR) placed bath, bedroom1, bedroom2, hallway, kitchen, living room, stairs, study room, and toilet.

2.2 Feature Extraction

Since we target to predict human activities within each second. We firstly divide the entire time series into 1-second-length segments. Based on each time segment, we extract useful features to discriminate activities.

Features from Accelerometer We generated totally 12 features from raw tri-axial accelerometer data, including *"kurtosis"* [2], *"approximate entropy"* [3], *"top-10 Frequency by FFT"* [4], *"FFT distribution kurtosis"*, *"average jerk"* [5], *"average absolute value"*, *"average value"* [4], *"median"*, *"standard deviation"* [4], *"maximum value"*, *"minimum value"*, *"maximum absolute value"*.

The above 12 features are selected because of their capability of distinguishing our 20 target human activities. For example, during exploring the data, we find that class *"a_jump"* gets much higher maximum acceleration than any other activities. For another instance, while most of the activities don't repeat periodically, there are still some activities which show obvious repetitive patterns(e.g. *"a_ascend"*, *"a_walk"*). With Fast Fourier Transform(FFT), it can be observed that their distributions of energy in frequency domain are apparently different from those who do not have periodic patterns (e.g. *"p_stand"*, *"a_bend"*). Most of these repetitive activities get relatively higher energy in low-frequency bands, while others get comparable energy in all frequency bands.

RGB-D Camera Sphere Data provides us the coordinates of camera bounding boxes, from which we can extract both 2D and 3D movement and shape features. The movement features are derived from the center of the bounding boxes. Based on the center movement, the average, standard deviation, gradient values are calculated as features. The shape information is sourced from the coordinates of the corners of bounding boxes. The width, length, and area of the 2D bounding boxes, and the width, height, length and volume of the 3D bounding boxes are computed, respectively. Based on these shape information, the average and standard deviation values are computed as features.

Location Feature Extraction The location information is from RSSI and PIR sensors. The average signal values of RSSI and PIR are computed as location information, while additionally, the standard deviation of RSSI within each second is also computed as an indicator of motion speed. These information will be used in predicting room occupancy for our hierarchical approach.

Temporal Activity Consistency Features There exists obvious temporal consistency of activities in our daily life, and taking this information into account as a prior knowledge usually helps activity recognition, especially for transition activities. For example, *"t_lie_sit"* usually happens in a period between *"p_lie"* and *"p_sit"*; *"t_bend'* usually follows *"p_bent"*. Some counter examples include that *"a_jump"* never happens with *"p_lie"*, and *"a_ascend"* never happens with *"p_squat"*, etc.
Hence, a preliminary classifier is built to get the preliminary prediction on each data observation. Based on this preliminary prediction, we then get the $(i-1)$-second prediction and $(i+1)$-second prediction as the temporal consistency feature for second i. Since there are 20 activities in the data set, there are totally 40 temporal consistency features(20 for $(i-1)$-second prediction, and 20 for $(i+1)$-second prediction)

2.3 Annotation Confidence Level

Since all activities are conducted with a series of continuous actions and labeled by multiple annotators, which suffers the issue of inter-annotator disagreement. The ratio of agreements are treated as confidence level, or probability score. We denote $\mathbf{P}_{ij}^{target}$ as the confidence score of an observation $i \in N$ assigned to class $j \in K$.

3 Methodology

In this paper, we describe a location based hierarchical classification on human activities. In our proposed hierarchy as shown in Figure 1, the top layer is a classifier which predicts the room occupancy of each data point, while the second layer consists of local leaf nodes classifier of each room.

3.1 Room Occupancy Classification

In this step, we are going to build a room occupancy model to assign each data point into totally 9 rooms (*bath, bed1, bed2, hall, kitchen, living, stairs, study, toilet*) based on the average values of RSSI and PIR signals within each second. Simple gradient boosting algorithm is employed. Multi-labeled result is allowed here, thus, ending up with matrix $\mathbf{P}_{train}^{room} \in \Re_{N_{train} \times R}$ and $\mathbf{P}_{test}^{room} \in \Re_{N_{test} \times R}$ for training and testing set respectively, where N is the number of data points, and $R = 9$ is the number of rooms. Hence, each data point $i \in N$ is assigned to room $r \in R$ with a confidence score $\mathbf{P}_{ir}^{room}$.

3.2 Leaf Nodes Classification

In this step, we aim to build a classifier for each leaf node (i.e. room). In a leaf node r, the data matrix $\mathbf{X}^r \in \Re_{N \times d}$ consists of all the data points which are assigned to this room r.

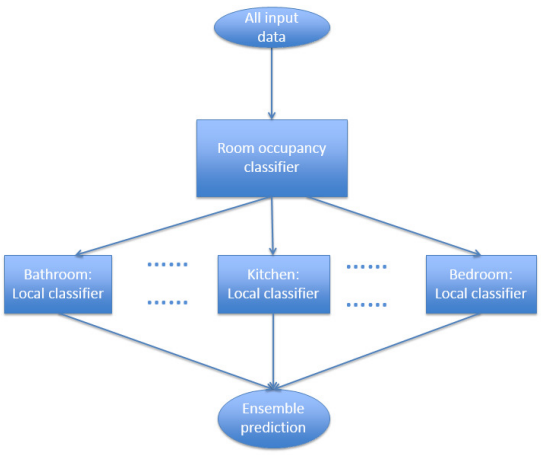

Fig. 1: The proposed hierarchy

Our approach employed gradient boosting algorithm and used a cost sensitive method for training on the leaf nodes. During the training phase, different data observations are treated with different costs which involves both the annotation confidence score $\mathbf{P}_{train}^{target}$ and the room occupancy confidence score $\mathbf{P}_{train}^{room}$. For training data point i, its training cost for class j on leaf node r is:

$$(\mathbf{W}_{train})_{ijr} = (\mathbf{P}_{train}^{room})_{ir} \times (\mathbf{P}_{train}^{target})_{ij}$$

The $(\mathbf{W}_{train})_{ijr}$ can be seen as a cost when training data point i is not assigned to class j by local model r.

During the prediction phase on testing data set, each local model on the leaf node r generates a matrix $(\mathbf{P}_{test}^{output})^r \in \Re_{N_{test} \times K}$, indicating the class assignment of the testing data set on K classes by the local model.

3.3 Ensemble Prediction

Since each leaf node r ends up with a prediction score matrix $(\mathbf{P}_{test}^{output})^r$, the following formula should be employed to ensemble these R prediction matrix into an ultimate matrix $(\mathbf{P}_{test}^{output}) \in \Re_{N_{test} \times K}$:

$$(\mathbf{P}_{test}^{output})_{ij} = \sum_{r}^{R} (\mathbf{P}_{train}^{room})_{ir} \times (\mathbf{P}_{test}^{output})_{ij}^{r}$$

4 Experimental Results

4.1 Evaluation Metric

The output of the algorithm is a prediction score matrix $\mathbf{P}^{output} \in \Re^{N \times K}$. Each element $\mathbf{P}_{ij}^{output}$ indicates the probability observation i is assigned to class j

by the algorithm. Similarly, the ground truth probability matrix is denoted as $\mathbf{P}^{target}$ which represents the confidence score of annotation.

Due to the class unbalance nature of the data, instead of accuracy, we report the experimental results with weighted brier score[6][1] and micro-f1 score [7].

4.2 Baselines

Several baselines are also tried for comparison. The simplest baseline is to use the prior class distribution for prediction, which is denoted as *"prior"*. We also tried the baseline feature set given in [8] which includes only the *"mean"*, *"median"*, *"std"*, *"max"* and *"min"* of the raw data and we denote this baseline as *"baseline feature"*. Finally, we tried our proposed feature set with/without temporal activity consistency, denoted as *"proposed feature"* and *"proposed feature+consistency"* repectively. In addition, we denote our ultimate approach as *"proposed feature+consistency+location hierarchy"*.

4.3 Experiment Results

Table 1: Performance of Each Method

methods	brier score	micro-f1 score
prior	0.256094	0.027281
baseline feature	0.17218	0.62683
proposed feature	0.154171	0.664165
proposed feature+location hierarchy	0.143692	0.689283
proposed feature+consistency	0.097795	0.757628
proposed feature+consistency+location hierarchy	0.095958	0.761101

The experimental results reported with 5-fold cross validation which are shown in Table 1. From this table, we can see that the weighted brier score of the proposed framework *"proposed feature+consistency+location hierarchy"* wins all the rest baselines. The *"proposed feature"* wins the *"baseline feature"*, indicating that our accelerometer and RGB-D camera features are useful in revealing the activity patterns. By introducing the temporal activity consistency features, the performance of *"proposed feature+consistency"* is enhanced a lot compared with *"proposed feature"*, which emphasizes again the importance of temporal consistency in inferring activities. The performance of *"proposed feature+location hierarchy"* also improves from *"proposed feature"* which means that our location based hierarchy helps.

By selecting the top-1 scored activity as the prediction for each observation, we obtain a single predicted label for each observation. The confusion matrix for *"proposed feature+consistency+location hierarchy"* is shown in Table 2.

5 Conclusion

In this paper, we describe a hierarchical structure for indoor human activity recognition by team **MSUHPDM** during ECML/PKDD 2016 Sphere Challenge. Useful accelerometer, RGB-D camera and location features are extracted.

Table 2: Confusion Matrix by *proposed feature+consistency+location hierarchy*

	a_ascend	a_descend	a_jump	a_loadwalk	a_walk	p_bent	p_kneel	p_lie	p_sit	p_squat	p_stand	t_bend	t_kneel_stand	t_lie_sit	t_sit_lie	t_sit_stand	t_stand_kneel	t_stand_sit	t_straighten	t_turn	precision (%)
a_ascend	**110**	0	0	1	8	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	90
a_descend	2	**104**	0	3	11	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	84
a_jump	0	0	**34**	0	1	0	0	0	0	0	5	0	0	0	0	0	0	0	0	1	83
a_loadwalk	7	1	0	**59**	7	0	0	0	0	0	7	0	0	0	0	0	0	0	0	2	71
a_walk	30	33	5	69	**1115**	15	0	0	8	0	254	19	2	0	0	5	1	4	20	199	63
p_bent	0	0	0	1	9	**453**	10	3	3	1	35	53	1	0	0	0	2	1	32	2	75
p_kneel	0	0	0	0	0	11	**207**	0	2	7	4	1	15	0	0	0	19	0	1	0	78
p_lie	0	0	0	0	2	10	4	**1238**	66	1	17	0	0	50	47	2	1	2	1	129	79
p_sit	0	0	2	0	31	21	4	90	**2738**	5	49	1	1	49	40	43	2	47	3	31	87
p_squat	0	0	0	0	0	3	16	0	0	**43**	4	1	1	0	0	0	1	0	0	0	62
p_stand	1	2	5	15	462	247	29	19	87	7	**5468**	87	43	4	3	57	32	54	89	366	77
t_bend	0	0	0	3	4	25	0	0	0	0	15	**66**	0	0	0	1	2	1	0	4	55
t_kneel_stand	0	0	0	0	2	1	15	0	2	2	12	0	**45**	0	0	1	0	0	0	0	56
t_lie_sit	0	0	0	0	2	0	0	13	21	0	0	0	1	**90**	0	1	0	0	0	4	68
t_sit_lie	0	0	0	0	0	0	0	24	24	0	1	0	0	2	**81**	0	0	0	1	0	61
t_sit_stand	0	0	0	1	5	0	1	0	21	0	16	1	1	0	0	**74**	0	0	1	1	61
t_stand_kneel	0	0	0	0	1	0	11	0	3	5	13	1	1	0	0	0	**64**	2	0	0	63
t_stand_sit	0	0	0	0	4	0	1	0	24	0	21	1	0	1	2	0	2	**103**	0	6	62
t_straighten	0	0	0	0	3	25	2	1	0	0	24	0	1	0	0	0	0	1	**75**	6	54
t_turn	0	0	0	4	38	6	2	10	2	0	46	7	0	4	2	1	0	7	3	**105**	44
recall (%)	73	74	74	38	65	55	69	89	91	61	91	28	40	45	46	40	51	46	33	12	76

The hierarchy is built based on the location information. In addition, temporal activity consistency is also involved in the framework. Our experimental result suggested the our proposed feature set is more effective than baseline features. Moreover, the location hierarchy and temporal activity consistency also helps a lot in recognizing human activities. Our proposed approach reaches the lowest weighted brier score as 0.0960 on 5-fold cross validation, with a micro-f1 score as 0.7611.

6 Related Work

Various sensors have been utilized to help elder people or patients who need special daily nursing, and one way to serve healthcare, is to track the daily activities of the patients. Among all varieties of sensors, inertial sensors are most popularly used, especially accelerometer and gyroscope. Many inertial sensors are also embedded into smart phones for more popular usage. In [9], a waist-mounted smart phone was worn by subjects to collect data on mainly 6 activities. In [10], miniature inertial and magnetic sensors were employed to classify on totally 19 human activities. All of above experiments fail to involve recognition on transition actions. Besides inertial sensors, more traditional activity detection with RGB cameras are also popular [11] [12]. In [13], a novel robot centric activity recognition system is proposed, which aims to recognize interaction activities. Additionally, passive environmental sensors like PIR are often helpful when building a smart environment [14] [15].

References

1. N. Twomey, T. Diethe, M. Kull, H. Song, M. Camplani, S. Hannuna, X. Fafoutis, N. Zhu, P. Woznowski, P. Flach, and I. Craddock, "The SPHERE challenge: Activity recognition with multimodal sensor data," *arXiv preprint arXiv:1603.00797*, 2016.
2. J. Baek, G. Lee, W. Park, and B.-J. Yun, "Accelerometer signal processing for user activity detection," in *Proceedings of International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, 2004, pp. 610–617.
3. S. M. Pincus, I. M. Gladstone, and R. A. Ehrenkranz, "A regularity statistic for medical data analysis," *Journal of clinical monitoring*, vol. 7, pp. 335–345, 1991.
4. L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *Proceedings of International Conference on Pervasive Computing*, 2004, pp. 1–17.
5. A. Weiss, T. Herman, M. Plotnik, M. Brozgol, N. Giladi, and J. Hausdorff, "An instrumented timed up and go: the added value of an accelerometer for identifying fall risk in idiopathic fallers," *Physiological measurement*, vol. 32, p. 2003, 2011.
6. G. W. Brier, "Verification of forecasts expressed in terms of probability," in *Monthly weather review*, 1950, pp. 1–3.
7. Y. Yang, "An evaluation of statistical approaches to text categorization," *Information Retrieval*, vol. 1, pp. 69–90, 1999.
8. "The sphere challenge: Activity recognition with multimodal sensor data," http://blog.drivendata.org/2016/06/06/sphere-benchmark/.
9. D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones." in *Proceedings of ESANN*, 2013.
10. K. Altun, B. Barshan, and O. Tunçel, "Comparative study on classifying human activities with miniature inertial and magnetic sensors," *Pattern Recognition*, vol. 43, pp. 3605–3620, 2010.
11. H. Zhang and L. E. Parker, "Code4d: color-depth local spatio-temporal features for human activity recognition from rgb-d videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, pp. 541–555, 2016.
12. A. Farooq, A. Jalal, and S. Kamal, "Dense rgb-d map-based human tracking and activity recognition using skin joints features and self-organizing map," *KSII Transactions on internet and information systems*, vol. 9, pp. 1856–1869, 2015.
13. L. Xia, I. Gori, J. K. Aggarwal, and M. S. Ryoo, "Robot-centric activity recognition from first-person rgb-d videos," in *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, 2015, pp. 357–364.
14. N. Twomey and P. Flach, "Context modulation of sensor data applied to activity recognition in smart homes," in *Proceedings of Workshop on Learning over Multiple Contexts, European Conference on Machine Learning (ECML'14)*, 2014.
15. M. Sathishkumar and S. Rajini, "Smart surveillance system using pir sensor network and gsm," *International Journal of Advnced Research in Computer Engineering & Technology*, vol. 4, 2015.