Towards Zero-Shot Learning for Human Activity Recognition Using Semantic Attribute Sequence Model

Heng-Tze Cheng, Martin Griss
Carnegie Mellon University
hengtze@cmu.edu, martin.griss@sv.cmu.edu

Paul Davis, Jianguo Li, Di You
Motorola Mobility
{pdavis, jianguo.li, di.you}@motorola.com

ABSTRACT

Understanding human activities is important for user-centric and context-aware applications. Previous studies showed promising results using various machine learning algorithms. However, most existing methods can only recognize the activities that were previously seen in the training data. In this paper, we present a new zero-shot learning framework for human activity recognition that can recognize an unseen new activity even when there are no training samples of that activity in the dataset. We propose a semantic attribute sequence model that takes into account both the hierarchical and sequential nature of activity data. Evaluation on datasets in two activity domains show that the proposed zero-shot learning approach achieves 70-75% precision and recall recognizing unseen new activities, and outperforms supervised learning with limited labeled data for the new classes.

Author Keywords

Activity recognition, zero-shot learning, semantic attributes.

ACM Classification Keywords

I.5.2 Pattern Recognition: Design Methodology—*Classifier design and evaluation*; C.3 Special-Purpose and Application-Based Systems: Real-time and embedded systems

INTRODUCTION

Human activity recognition is an important element that enables many context-aware applications in the area of pervasive and ubiquitous computing [8]. There has been extensive study on activity recognition using supervised learning methods, where a classifier is trained on a large set of labeled examples of every target activity (see Figure 1) [3, 8]. While many promising results have been reported, a widely acknowledged problem is that labeled examples are often time consuming and expensive to obtain, as they require a lot of effort from test subjects or domain experts [14, 15]. Semi-supervised learning has been proposed to improve recognition accuracy on the seen activities by leveraging unlabeled data [15]. However, most existing methods still cannot recognize a previously unseen new activity if there were no training samples of that activity in the dataset.

In light of these limitations, the research question we aim to answer is: *Given a sequence of sensor data, how to recognize a human activity even when no training examples of that activity are available?* This is often referred to as the *zero-shot learning* problem, where the goal is to learn a classifier that can recognize new classes that are not in the training data. Recently, NuActiv [5] presented an early study on zero-shot learning for activity recognition. However, NuActiv classifies each frame independently and does not take into account the temporal dependency of the semantic attributes. Also, it has not been shown that how zero-shot learning compares to supervised learning for activity recognition.

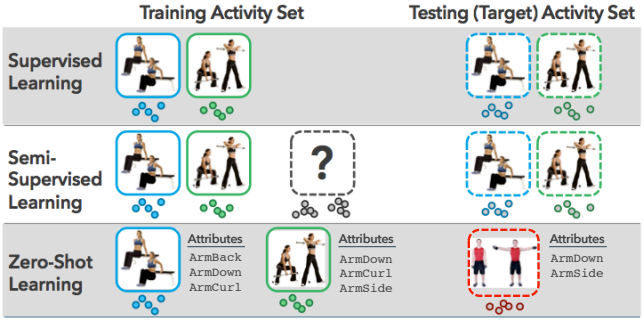


Figure 1. Different problem settings and learning methods for activity recognition. Supervised learning only learns from labeled samples of the target activities. Semi-supervised learning leverages additional unlabeled samples. The proposed zero-shot learning method generalizes learned mid-level attributes to recognize unseen new activities.

In this paper, we extend the previous work and tackle the problem using two ideas. First, an unseen activity may possess some underlying semantic attributes that can be found in some seen activities, where each of the attributes is a human readable term that describes a basic element or an intrinsic characteristic of an activity. For example, the attributes "`Sitting`" and "`HandsOnTable`" can be observed in of both "having lunch" and "working at desk" activities. Second, the sequence of human activities and the underlying attributes are time series with strong temporal dependency. Integrating these ideas, we propose a new zero-shot learning framework for activity recognition. For experiments, we use the same attributes and dataset as those used in [5]. The main contributions of the paper are: (1) The design and implementation of the new *semantic attribute sequence* model that learns the hierarchical and sequential nature of human activities, and can recognize an activity even when there are no training data for that activity. (2) The evaluation of the approach in two real-world activity domains, and the comparison with supervised learning methods.

RELATED WORK

Most existing work in activity recognition addresses the seen-class recognition problem using supervised learning [3, 8] or

UbiComp'13, September 8–12, 2013, Zurich, Switzerland.

10.1145/2493432.2493511

Table 1. List of notations used in the graphical model.

Symbol	Description
Y_t	High-level activity class label at time t
A_t	Mid-level semantic attribute at time t
$\mathbf{x}_t$	Low-level D-dimensional feature vector at time t. $\mathbf{x}_t = \{X_{1,t}, X_{2,t}, ..., X_{D,t}\}$
$\phi_k(\mathbb{V})$	The k-th potential function in a probabilistic graphical model involving a set of vertices $\mathbb{V}$

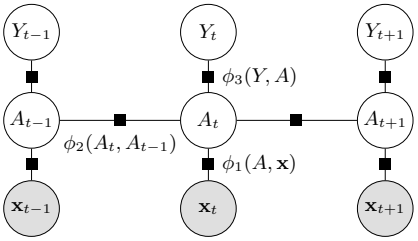


Figure 2. Graphical model of sequence of high-level human activities, mid-level semantic attributes, and observed low-level signal features.

semi-supervised learning [14, 15]. Transfer learning has also been studied where the instances or models for activities in one domain can be transferred to improve the recognition accuracy in another domain and reduce the need of training data [2, 17]. Our work extends the previous work to address the zero-shot learning problem. Another direction in this area is unsupervised learning, which focuses on clustering or pattern discovery [7]. Although labels are not required for unsupervised learning, the output is a set of unnamed clusters, which cannot be used for recognition purposes.

The concept of zero-shot learning and attributes has shown promise in the field of object recognition [9, 13], neural activity recognition [12], and human action recognition from videos [6, 10, 16]. However, the visual attributes cannot be directly applied to activity recognition using sensor data from mobile devices. Our work extends the previous work by proposing a new zero-shot learning framework for sequential sensor data using a graphical model and comparing it to supervised learning with limited training data.

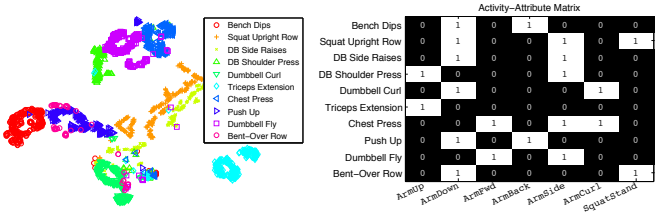
PROBLEM DEFINITION

Consider a set of target human activity classes, $\mathbb{Y}$, that we aim to recognize. $\mathbb{Y} = \{\{y_1, y_2, ..., y_s\}, \{y_{s+1}, ..., y_{s+u}\}\} = \mathbb{Y}_S \cup \mathbb{Y}_U$. $\mathbb{Y}_S$ is the set of *seen activity classes*, where there exists some training data for every class. $\mathbb{Y}_U$ is the *unseen activity classes* set where there are no training data for any class. The problem is: How to train a model to recognize an unseen activity class $y \in \mathbb{Y}_U$ given a set of N training instances $\{(\mathbf{x}^{(i)}_{train}, y^{(i)}_{train})\}_{i=1}^{N}$, each including a feature vector $\mathbf{x}^{(i)}_{train}$ and a ground truth class label $y^{(i)}_{train} \in \mathbb{Y}_S \; \forall \; i$?

PROPOSED MODEL AND METHOD

Graphical Model of Semantic Attribute Sequences

To tackle the problem, we designed a probabilistic graphical model in Figure 2 to recognize seen or unseen activities through a layer of semantic attributes (the notations are listed in Table 1). The model is a variation of conditional random field (CRF) [1], which is suitable for activity recognition because it models the temporal dependency in sequential data [8]. It also supports the use of complex features, whose distributions and dependencies may not have a simple parametric form, by imposing weaker assumptions on the dependencies



(a) Features of activity samples. (b) Activity-Attribute Matrix.

Figure 3. Activities in the (a) feature space and (b) attribute space.

between features compared to hidden Markov models [1, 2]. Given a sequence of observed features $\boldsymbol{X} = \{\mathbf{x}_1, ..., \mathbf{x}_T\}$, the conditional probability distribution of the activity sequence $\boldsymbol{Y} = \{Y_1, ..., Y_T\}$, the attribute sequence $\boldsymbol{A} = \{A_1, ..., A_T\}$ given $\boldsymbol{X}$ is modeled as:

$$P(\boldsymbol{Y}, \boldsymbol{A}|\boldsymbol{X}) = \frac{1}{Z(\boldsymbol{X})} \prod_{t=1}^{T} \prod_{k=1}^{K} \phi_{k,t}(Y_t, A_t, A_{t-1}, \mathbf{x}_t) \quad (1)$$

where $Z(\boldsymbol{X}) = \sum_{\boldsymbol{Y}} \sum_{\boldsymbol{A}} \prod_{t=1}^{T} \prod_{k=1}^{K} \phi_{k,t}(Y_t, A_t, A_{t-1}, \mathbf{x}_t)$ is a normalization term that ensures the probability distribution sums up to one. Each $\phi_{k,t} = \omega_k f_k(Y_t, A_t, A_{t-1}, \mathbf{x}_t)$ is a potential function that consists of a model parameter ω_k and a feature function f_k defined over a subset of the random variables Y, A, and $\mathbf{x}$. Our probabilistic graphical model consists of three types of potential functions:

- $\phi_{1,t}$ models the probability distribution of an attribute A_t given a feature vector $\mathbf{x}_t$:

$$\phi_{1,t}(A_t, \mathbf{x}_t) = \exp(\sum_{a \in \mathbb{A}} \sum_{d=1}^{D} \omega_{a,d} x_{d,t} \cdot I(A_t = a)) \quad (2)$$

where $I(p)$ is the indicator function that takes the value 1 if the statement p is true, and takes the value 0 otherwise.

- $\phi_{2,t}$ models the temporal dependency between neighboring semantic attribute values:

$$\phi_{2,t}(A_t, A_{t-1}) = \exp(\sum_{a \in \mathbb{A}} \sum_{a' \in \mathbb{A}} \omega_{a,a'} I(A_t = a) I(A_{t-1} = a')) \quad (3)$$

- $\phi_{3,t}$ models the correlation between the activity class Y and the semantic attribute A:

$$\phi_{3,t}(Y_t, A_t) = \exp(\sum_{y \in \mathbb{Y}} \sum_{a \in \mathbb{A}} \omega_{y,a} I(Y_t = y) I(A_t = a)) \quad (4)$$

Low-Level Feature Extraction

The bottom part of the graphical model is the sequence of low-level signal features. Our model does not make assumptions on the input data type or the generative distributions of the features so any kind of sensor data can be used. In this work we use inertial sensor data as an example. The sensor data stream is first segmented using overlapping windows (see parameters in the Evaluation section). For each segment the following features are computed: The *mean* and *standard deviation*, *pairwise correlation* between each pair of dimensions, *local slope* of the sensor data fitted by 1st-order linear regression, and *zero-crossing rate*. All features are computed for each dimension, x, y, and z. To visualize the effectiveness of the features, we plot the dataset in the feature space in Figure 3(a) (dimension reduced to 2 for visualization using t-SNE [11]), where the samples form natural clusters with minor overlaps. For the daily life activity dataset [7], we also include *time of day* as an input feature, as it is correlated with the daily life routines of a user.

Mid-Level Semantic Attributes

The relationship between mid-level semantic attributes and high-level activities is encoded in an Activity-Attribute Matrix as shown in Figure 3(b). Each element m_{ij} represents the association between activity i and attribute j. Currently we use binary values, indicating whether such an association exist ($m_{ij} = 1$) or not ($m_{ij} = 0$). In general, m_{ij} can be real-valued ($0 \leq m_{ij} \leq 1$), indicating the confidence or level of the association. Activity-Attribute Matrix can be thought as an auxiliary input that substitute for labeled sensor data by encoding the correlation between the attributes and the seen/unseen activities (i.e. $\phi_{3,t}$). For the current implementation, the activity-attribute matrix is manually defined by common-sense knowledge and domain knowledge as an initial attempt towards zero-shot learning. Our definition is inspired by the attributes defined in [10]. It is to be noted that automating the process is possible using web text mining, as explored in zero-shot learning literature [13]. A user can also provide a one-time definition of a custom new activity by simply describing it using the semantic attributes, which is equivalent to inserting a row into the matrix.

Activity Recognition through Sequence Decoding

The goal of activity recognition is to decode the sequence of activities $\boldsymbol{Y}$ (may be seen or unseen) given a sequence of observed features $\boldsymbol{X}$, through a layer of attribute sequence $\boldsymbol{A}$. During the offline training phase, the optimal model parameters $\theta^* = \{\omega_k^*\}$ are learned by maximizing the regularized log-likelihood of the training data $L(\theta)$:

$$L(\theta) = \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{k=1}^{K}\omega_k f_k(Y_t^{(i)}, A_t^{(i)}, A_{t-1}^{(i)}, \mathbf{x}_t^{(i)}) - \sum_{i=1}^{N}\log Z(\boldsymbol{X}^{(i)}) - \sum_{k=1}^{K}\lambda\omega_k^2 \quad (5)$$

where the last term is the L^2-regularization term that penalizes large ω_k values with a weighting λ to prevent overfitting [1]. λ is empirically set to 50 based on a cross-validation test. The optimization problem $\theta^* = \text{argmax}_\theta\, L(\theta)$ is solved using L-BFGS [1], a widely used optimization algorithm. During the online testing phase, the states of the sequence $\boldsymbol{Y}$, $\boldsymbol{A}$ with maximum likelihood are decoded using the Junction Tree algorithm [1]. Since all the target class in $\mathbb{Y}$ are described using the semantic attributes in the attribute space $\mathbb{A}$ in the form of an Activity-Attribute Matrix, we are able to decode the value y_t^* that Y_t takes on even if y_t^* corresponds to an previously unseen new activity, i.e. $y_t^* \in \mathbb{Y}_U$.

EVALUATION

Exercise Activity Dataset

We used the exercise activity dataset in [5]. 20 test subjects were asked to perform a set of 10 exercise activities as listed in Figure 3(b) with 10 iterations. Descriptions of these activities can be found in [4]. Each subject was equipped with three sensor-enabled devices: A Nexus S 4G phone in an armband, a MotoACTV wristwatch, and a second MotoACTV clipped to the hip. The dataset contains accelerometer and gyroscope data collected at 30 Hz sampling rate. For feature extraction, the sliding window size is empirically set to 1 second with 50% overlap based on a cross-validation test.

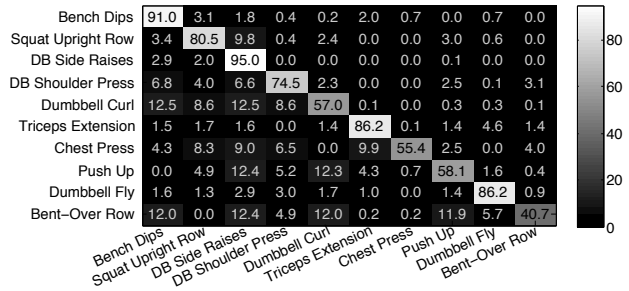


Figure 4. Confusion matrix of unseen exercise activity recognition in percentages (rows: ground-truth classes, columns: estimated classes).

Daily-Life Activity Dataset

For daily life activities, we used the dataset from TU Darmstadt [7, 14]. The dataset includes 34 daily life activities collected for 7 days. The data were collected with an accelerometer worn on the wrist and hip of the subject. The sampling rate is 100Hz, and the features are computed from a sliding window of 30 seconds with 50% overlap. We used the same attribute list and activity-attibute matrix as those used in [5].

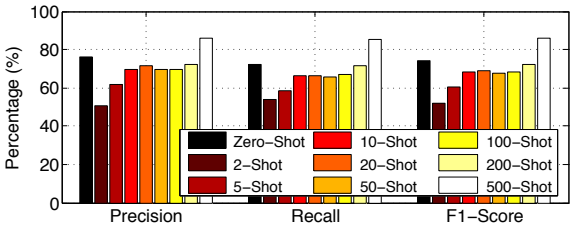
Evaluation Methodology

We used leave-two-class-out cross validation, the most widely used validation method used in the literature of zero-shot learning [5, 12]. The validation scheme is used for recognizing unseen classes that do not have any sample in the training set. For a total of N classes, we first train our system on $(N-2)$ classes, and then test the classifier on the remaining 2 classes that were "unseen" to the system during training. We repeat the test for all $\binom{N}{2}$ combinations and report the average results in precision (P), the percentage of times that a recognition made by the system is correct, and recall (R), the percentage of times that an activity done by a user is detected by the system. F_1-score $= (2PR/(P+R))$ is an integrated measure that combines both [3].

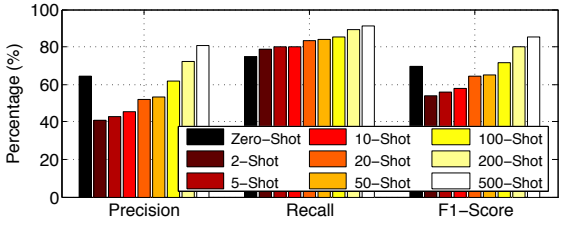
Case Study I: Unseen Exercise Activity Recognition

As shown in the confusion matrix in Figure 4, our approach achieved 76% precision and 72% recall averaged over all activities. The standard deviation of precision and recall across different users is 13% and 16%, respectively. The results show that even without training samples, unseen new activities can be recognized with a reasonable accuracy. The results are also comparable to those reported in [5]. On the other hand, the limitation of the approach is observed when two activities only differ in one or two attributes (e.g. Push Up and Dumbbell Side Raises), or when an attribute is not consistent for every person (e.g. `ArmCurl`).

We compare our approach with existing supervised learning approach using a linear-chain CRF [8], which belongs to the same model family except without the semantic attribute layer. Since supervised learning cannot recognize unseen activities without training samples, we compare with the cases where it is possible to obtain n samples of each unseen activity performed and labeled by the users (denoted by n-shot learning). In contrast, zero-shot learning can be thought as having users provide a one-time description of an unseen activity using the semantic attributes. As shown in Figure 5(a), zero-shot learning outperforms supervised learning with up to



(a) Exercise activities.



(b) Daily life activities.

Figure 5. Comparison between proposed zero-shot learning and n-shot supervised learning (n labeled training samples for each target activity).

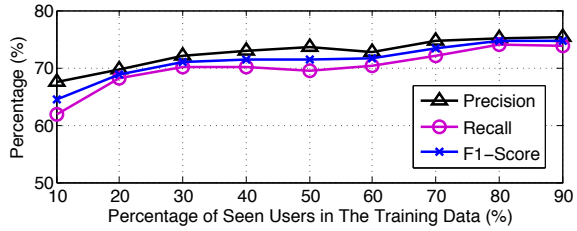


Figure 6. Cross-user recognition accuracy. The testing set includes 10% of the users, which are different from those in the training data.

200 labeled samples. This shows that zero-shot learning is effective for bootstrapping an activity recognition system when sufficient labeled samples for every activity are not available. On the other hand, in cases where obtaining a large amount of labeled data is inexpensive for every activity, supervised learning tends to achieve higher accuracy.

Case Study II: Unseen Daily Life Activity Recognition

We applied the same approach to the daily life activity dataset. The average precision and recall is 69% and 75%, respectively, which outperformed the results (52.3% precision and 73.4% recall) reported in [5]. The results suggest that the semantic attribute sequence model can better capture the temporal dependency in the daily life activities, in comparison to applying an activity classifier to each frame independently [5]. As shown in Figure 5(b), the F_1-score of zero-shot learning outperforms supervised learning approach with less than 100 labeled samples. Possible reasons for a lower precision include a larger number of different activity classes, and a larger feature variation in daily life activities because they are less well-defined than exercise activities.

Cross-User Unseen Activity Recognition Results

The ability to apply the model learned from some users to new users is important for an activity recognition system. Figure 6 shows the results on unseen exercise activity recognition when the users in the training set and testing set are different. Our approach achieves a stable accuracy of 70-75% when 20% or more users were seen in the training set. The results show the learned attribute sequence model can be generalized to new users and new activities, and the performance degrades gracefully with the decrease of seen users. In comparison, in the experiments where the training data and testing data are drawn from the same user, the average precision and recall increases to 78% and 76%, respectively.

CONCLUSION

In this paper, we have presented the design, implementation, and evaluation of a new zero-shot learning framework for human activity recognition. Most existing activity recognition systems cannot recognize a previously unseen new activity if there were no training samples of that activity in the dataset. The proposed semantic attribute sequence model learns the relationship between activities and features through a semantic attribute layer. This enables the reuse and generalization of learned attribute models for recognizing unseen new activities. Evaluation results show that our approach achieves 70-75% precision and recall in unseen activity recognition, and outperforms supervised learning with an order of hundreds of labeled data for the new classes.

REFERENCES

1. Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer-Verlag Inc., 2006.
2. Blanke, U., and Schiele, B. Remember and transfer what you have learned - recognizing composite activities based on activity spotting. In *Int'l Symp. Wearable Computers* (2010).
3. Cao, H., Nguyen, M. N., Phua, C., Krishnaswamy, S., and Li, X.-L. An integrated framework for human activity classification. In *Int'l Conf. Ubiquitous Computing* (2012).
4. Chang, K.-H., Chen, M. Y., and Canny, J. Tracking free-weight exercises. In *Proc. Int'l Conf. Ubiquitous computing* (2007).
5. Cheng, H.-T., Sun, F.-T., Griss, M., Davis, P., Li, J., and You, D. NuActiv: Recognizing unseen new activities using semantic attribute-based learning. In *Proc. Int'l Conf. Mobile systems, applications, and services*, MobiSys '13 (2013), 361–374.
6. Fu, Y., Hospedales, T. M., Xiang, T., and Gong, S. Attribute learning for understanding unstructured social activity. In *Proc. European Conf. Computer Vision* (2012), 530–543.
7. Huynh, T., Fritz, M., and Schiele, B. Discovery of activity patterns using topic models. In *Proc. Int'l. Conf. Ubiquitous Computing*, UbiComp '08 (2008), 10–19.
8. Kim, E., Helal, S., and Cook, D. Human activity recognition and pattern discovery. *IEEE Pervasive Computing* (2010).
9. Lampert, C. H., Nickisch, H., and Harmeling, S. Learning to detect unseen object classes by between-class attribute transfer. In *Int'l Conf. Comp. Vision and Patt. Recog.* (2009).
10. Liu, J., Kuipers, B., and Savarese, S. Recognizing human actions by attributes. In *Conf. Computer Vision and Pattern Recognition* (2011), 3337–3344.
11. Maaten, L., and Hinton, G. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research* (2008).
12. Palatucci, M., Pomerleau, D., Hinton, G. E., and Mitchell, T. M. Zero-shot learning with semantic output codes. In *Proc. Neural Information Processing Systems* (2009).
13. Parikh, D., and Grauman, K. Interactively building a discriminative vocabulary of nameable attributes. In *Proc. Int'l Conf. Computer Vision and Pattern Recognition* (2011).
14. Stikic, M., Larlus, D., Ebert, S., and Schiele, B. Weakly supervised recognition of daily life activities with wearable sensors. *IEEE Trans. PAMI* (2011).
15. Stikic, M., Van Laerhoven, K., and Schiele, B. Exploring semi-supervised and active learning for activity recognition. In *Int'l Symp. Wearable Computers* (2008), 81–88.
16. Yao, B., Jiang, X., Khosla, A., Lin, A. L., Guibas, L., and Fei-Fei, L. Human action recognition by learning bases of action attributes and parts. In *Proc. ICCV* (2011).
17. Zheng, V., Hu, H., and Yang, Q. Cross-domain activity recognition. In *Int'l Conf. Ubiquitous Computing* (2009).