

---

# Learning Human Activities and Object Functionalities through Conditional Random Fields

---

**Nakul Gopalan**  
Department of Computer Science  
Brown University  
Providence, RI 02912  
ngopalan@cs.brown.edu

**Jun Ki Lee**  
Department of Computer Science  
Brown University  
Providence, RI 02912  
jkleee@cs.brown.edu

## Abstract

Recognizing a human activity is a skill that humans learn early in life. Activity recognition helps humans in predicting next steps of the activity in question and also in planning other activities with respect to the recognized activity. Our goal is to recognize human activities using machine learning, to predict next steps of an action, or provide help using robots or other interfaces. Specifically, what interests us is the human activities being performed near a table-top table or a shelf, where the human can be observed and helped by a robot. We will use a dataset from [2] to learn a Graphical Model with temporal and spatial structure, which can then be used to predict the action being performed and functionalities of objects. Both will then be used to recognize a higher level activity. As a baseline we have results of a bag of words classification of the 10 activities, which yields a result of 67%. We will tackle the problem using Conditional Random Fields and try to achieve better performance.

## 1 Introduction

Recognizing a human activity is a skill that humans learn early in life. Activity recognition helps humans in predicting next steps of the activity in question and also in planning other activities with respect to the recognized activity. Our objective is to recognize human activities using machine learning, to predict next steps of an action. Recognized actions can provide help to humans using robots or other interfaces. Specifically of interest to us are the human activities being performed on a table-top, where the human can be observed and helped by a robot.

We plan to use a dataset<sup>1</sup> used in [2] to learn a Graphical Model with temporal and spatial structure, which can then be used to predict the action being performed and functionalities of objects. Both will then be used to recognize higher level activities. This dataset is created from streams of the RGB-D images acquired from a Infrared depth sensing camera. The dataset contains both the tracked human skeletal poses and object locations and supervised labels of sub- and high-level- activities and object functionalities.

Among many works in human activity recognition using RGB-D images, Sung et al. in [6] proposed a hierarchical maximum entropy Markov model to detect high level activities and sub-activities. This work only uses human skeletal data. Gall et al. in [1] used both the human skeletal data and object tracking data to inference sub-activities and object functionalities. His approach sequentially inferences sub-activities and then object functionalities. It does not take both into account at the same time. Koppula et al. in [2] and [3], uses both the human skeletal poses and object locations and their relationships to low and high level activities and functionalities of objects using Conditional Random Fields (CRF).

---

<sup>1</sup><http://pr.cs.cornell.edu/humanactivities>

## 2 Dataset

We explored data sets with activity labels and corresponding skeletal trajectories. Humans frequently perform activities using objects, especially in table-top scenarios. The dataset from [2] was constructed keeping this relationship in mind. The raw data consists of the RGB-D frames from the scenes of humans performing activities and human skeletal poses from the OpenNI NITE tracker [4]. The skeletal poses are not precise and have confidence intervals. The labels on the data are related to both the objects present in the scene and the activities involved. Activity specific labels are framewise labels on subactivities and the main activity being performed. The sub-activity labels are that of *reaching*, *moving*, *pouring*, *eating*, *drinking*, *opening*, *placing*, *closing*, *scrubbing* and *null*. The object specific data consist of object locations for each time frame per object with its object id and its affordance present in the scene. The object affordance labels are those of *reachable*, *movable*, *pourable*, *pour to*, *containable*, *drinkable*, *openable*, *placeable*, *closable*, *scrubbable*, *scrubber* and *stationary*. The dataset also has precomputed features that have relationship between objects and skeletal poses across frames. For the experiments performed in this proposal we use only skeleton pose data, but for our next set of results we would use the pairwise relationship between objects and the skeleton data, and also the temporal features of data available.

## 3 Methods

The data from [2] has both temporal and spatial structure. For example, the spatial structure has relationships between objects and skeleton positions within a frame. There is also a temporal relationship between previous skeletal pose and object positions and those of the current frame. The authors from [2] modelled this structure using a pairwise Conditional Random Field([7]) as shown in Fig. 1a. Here each node corresponds to an object or a subactivity. The object nodes can take any of the 12 possible affordance labels and the subactivity nodes can take any of the 10 possible labels. The affordance and subactivity nodes are hidden, but feature values for label assignments over the nodes can be computed. The authors in [2] first predicted sub-activity labels and object affordances, i.e. inference over the model, and then used an SVM to predict the label of the over all activity. To predict the sub-activity and object affordance labels the authors used a maximum energy formulation, where the labels that lead to the maximum possible energy together are accepted. This energy is a sum of products of features and feature weights. Hence we need to learn the weights of the features first, which is performed in the learning stage, with the supervised labelled input from the dataset discussed above, such that training error is minimum. We plan to run the complete experiments ourselves with different features that the ones the authors used, or create some more. Then perform the learning and inference, and compare our results to those from the original paper.

As a baseline for our experiments we classify high level activities from skeletal trajectories using a Bag of Words model [5] with skeletal poses as features. We have about 8 examples per activity for training and 4 examples per activity for testing. Each frame has a skeletal pose, of 170 dimensions, and each activity has about 400 frames. We cluster skeletal poses over all activities into 200 clusters using k-means. Now for each activity label we create a histogram by counting clusters closest to each of the poses, and each activity has a histogram word. Now for each test activity we find the nearest neighbour word from the trained histogram words and return its label. The accuracy of this simple Bag of Words model is about 67.3%. In the next section we have attached a schedule that we will follow over the course of the project.

## 4 Schedule

**Week 0** (~11/11/14) Understanding the data set and running bag-of-words model classifier on the raw data set to classify high-level actions.

**Week 1** (~11/18/14) Extracting features from the raw data set.

**Week 2** (~11/25/14) Modelling Conditional Random Fields (CRFs)

**Week 3** (~12/02/14) Implementing the CRFs and Implementing Optimization Techniques

**Week 4** (~12/09/14) Pulling out the final result and Presentation Preparation

**Week 5** (~12/16/14) Pulling out the final result and write ups.

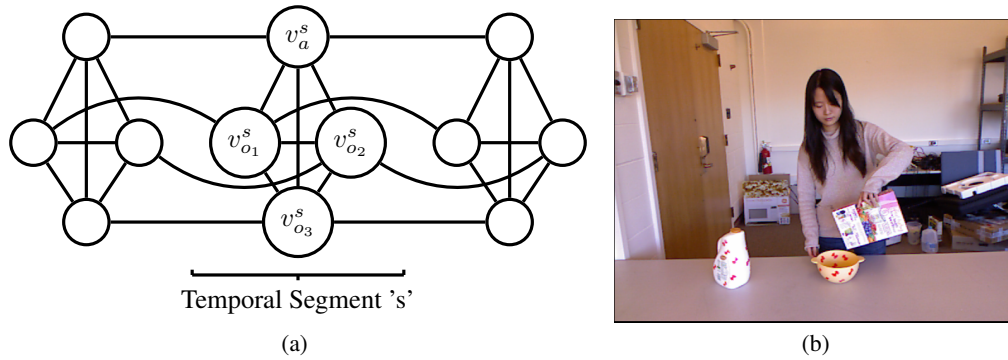


Figure 1: **(a)** an example of a Conditional Random Field used in [2]. Each temporal segment (s) has one sub-activity node  $v_a^s$  and three object nodes  $\{v_{o_1}^s, v_{o_2}^s, v_{o_3}^s\}$ , **(b)** an example image from the data set

## References

- [1] Juergen Gall, Andrea Fossati, and Luc Van Gool. Functional categorization of objects using real-time markerless motion capture. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1969–1976. IEEE, 2011.
- [2] Hema Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *IJRR*, 2013.
- [3] Hema Swetha Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. In *Robotics: Science and Systems*, 2013.
- [4] PrimeSense Inc. *Prime Sensor NITE 1.3 Algorithms notes*, 2010.
- [5] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, 2003.
- [6] Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. Unstructured human activity detection from rgb-d images. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 842–849. IEEE, 2012.
- [7] Charles Sutton and Andrew McCallum. *Introduction to Conditional Random Fields for Relational Learning*. MIT Press, 2006.