# 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
jklee@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 activitiy 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, or provide help using robots or other interfaces, or help recognizing problems in the activity being performed. 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 from [1] to learn a Graphical Model with temporal and spatial structure, which can then be used to predict the action being performed. As a baseline we have results of a bag of words classification of the 10 activities, which yields a result of 67%.

## 1 Introduction

## 2 Background

## 3 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 [1] 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 [2]. 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 labels consist of object location, object pose, object id and object affordance of the objects present in the scene. The object affordance lables 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.

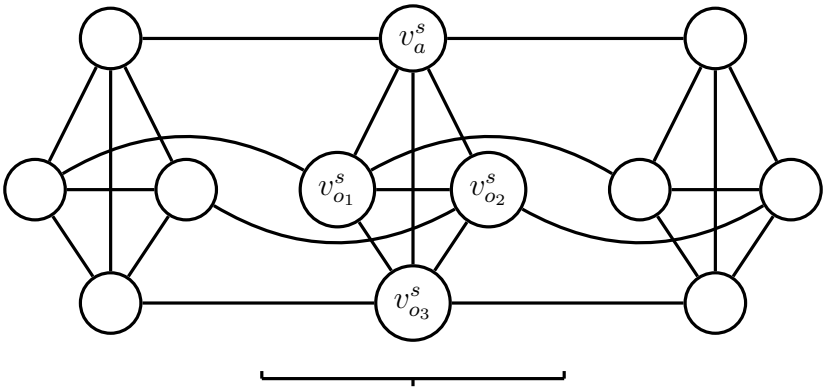


Figure 1: MRF

## 4 Methods

The data from [1] has a lot of structure, temporal and within the frame relationships between objects and skeleton positions. The authors from [1] modelled this structure using a pairwise graphical model as shown in Fig. 1. As a baseline for our experiments we create a try to classify skeletal trajectories using a Bag of Words model [3] with skeletal poses as features.

## 5 Schedule

**Week 0** ($\sim$11/11/14) Understanding the basic data set and running bag-of-worlds model classifier on the data set to classify high-level actions.

**Week 1** ($\sim$11/18/14) Extracting features from the raw data.

**Week 2** ($\sim$11/25/14) Applying basic probabilistic graphical models such as Markov Random Fields (MRFs).

**Week 3** ($\sim$12/02/14) Applying Conditional Random Fields (CRFs) to improve the performance.

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

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

## References

[1] Hema Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *IJRR*, 2013.

[2] PrimeSense Inc. *Prime Sensor NITE 1.3 Algorithms notes*, 2010.

[3] 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.