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 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 can then be used to recognize a higher level activity. First we compare the graph structures used in the original paper to understand the significance of the graph structure used. Further, our efforts in this work are to model the problem as an HMM first and then a CRF and compare performance to the original paper in classifying subactivities. We classify the main activity using a bag of words model with features from the complete activity.

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 use the dataset[1] 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. The 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. We first provide some background over the methods used to solve this problem before and an overview of the dataset. Next we alter the original graph structure in [2] by changing their code and running learning and testing on the new graph structure. Finally we try to predict the subactivities using linear chain HMMs and CRFs.

[1] http://pr.cs.cornell.edu/humanactitivities

2 Previous work

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).

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 [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 dataset has recordings of 120 labelled moderate sized human activities of ten types, e.g.: making cereal, microwaving, taking food, etc. 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 subactivitiesand 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.

4 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 affordace 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 segmented an input activity, or set of frames into segments that are more likely to belong to a single subactivity. The authors used a sum of Eucledian distance between joints metric to judge the change in sub-activities. The threshold to split subactivities seems to be based on a trial and error heuristic. The sub-activities are temorally linked giving a chain graph structure as shown in Fig. 1a. The authors used this graph structure to predict sub-activity labels and object affordances by performing 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. The energy function requires learning the weights of the features first, with the supervised labelled input from the dataset discussed above, such that training error is minimized.

5 Importance of graph structures

The original paper by [2] uses the graph structure shown in Fig. 1a. The subactivity nodes and object affordance nodes are completely connected within a temporal segment. The authors describe an energy function for the linear chain that describes the sequence of subactivities:

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}}\, E_{\mathbf{w}}(\mathbf{x}, \mathbf{y}) \tag{1}$$

$$E_{\mathbf{w}}(\mathbf{x}, \mathbf{y}) = E_o + E_a + E_{oo} + E_{oa} + E_{oo}^t + E_{oa}^t \tag{2}$$

$$E_o = \sum_{i \in \mathcal{V}_o} \psi_o(i) = \sum_{i \in \mathcal{V}_o} \sum_{k \in K_o} y_i^k \left[\mathbf{w}_{\mathbf{o}}^k \cdot \phi_o(i)\right] \tag{3}$$

$$E_a = \sum_{i \in \mathcal{V}_a} \psi_a(i) = \sum_{i \in \mathcal{V}_a} \sum_{k \in K_a} y_i^k \left[\mathbf{w}_{\mathbf{a}}^k \cdot \phi_a(i)\right] \tag{4}$$

$$E_{oo} = \sum_{(i,j) \in \mathcal{E}_{oo}} \sum_{(l,k) \in K_o \times K_o} y_i^l y_j^k \left[\mathbf{w}_{\mathbf{oo}}^{lk} \cdot \phi_{oo}(i,j)\right] \tag{5}$$

$$E_{oa} = \sum_{(i,j) \in \mathcal{E}_{oa}} \sum_{(l,k) \in K_o \times K_a} y_i^l y_j^k \left[\mathbf{w}_{\mathbf{oa}}^{lk} \cdot \phi_{oa}(i,j)\right] \tag{6}$$

$$E_{oo}^t = \sum_{(i,j) \in \mathcal{E}_{oo}^t} \sum_{(l,k) \in K_o \times K_o} y_i^l y_j^k \left[\mathbf{w}_{\mathbf{oo}}^{t^{lk}} \cdot \phi_{oo}^t(i,j)\right] \tag{7}$$

$$E_{aa}^t = \sum_{(i,j) \in \mathcal{E}_{aa}^t} \sum_{(l,k) \in K_a \times K_a} y_i^l y_j^k \left[\mathbf{w}_{\mathbf{aa}}^{t^{lk}} \cdot \phi_{aa}^t(i,j)\right] \tag{8}$$

where $E_w(x, y)$ is the total energy over all the temporal segments in the activity, x are the features and y are the subactivity and object affordance labels. V_a and V_o are the sub-activity and object nodes of the graph. K_a and K_o are the sub-activity and object affordance labels that the nodes can take. The energy functions are over individual nodes and edges of the graph, i.e., E_o and E_a are sums over the nodes taking on the labels y^k with features ϕ, when the weights of the nodes for label k are The individual energy terms of E_{oo}, E_{oa}, E_{oo}^t and E_{aa}^t $\phi(i)$ are the features

We plan to run the complete experiments ourselves with different features that the ones the authors used, or create some 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.

6 Schedule

Week 0 (∼11/11/14) Understanding the data set and running bag-of-worlds 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)

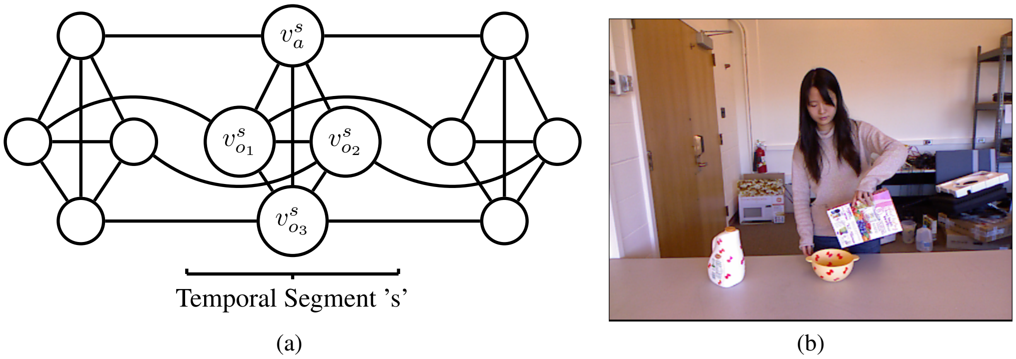
Week 3 (∼12/02/14) Implementing the CRFs and Implementing Optimization Techniques

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.

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.



(a)

(b)

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

[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 rgbd 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.