Predicting Human Activities Using Stochastic Grammar

Siyuan Qi\textsuperscript{1}, Siyuan Huang\textsuperscript{1}, Ping Wei\textsuperscript{2,1}, and Song-Chun Zhu\textsuperscript{1}

\textsuperscript{1}University of California, Los Angeles, USA
\textsuperscript{2}Xi’an Jiaotong University, Xi’an, China

\{syqi, huangsiyuan\}@ucla.edu pingwei@xjtu.edu.cn sczhu@stat.ucla.edu

1. Introduction

Consider the image from a video shown in Figure 1(a). A modern computer vision algorithm might reliably detect a human pose and some key objects in the scene: a chair, a monitor, a cup, a microwave and a water cooler. However, we as observers are able to reason beyond the current situation. We can predict what the possible future states are to some extent, and we can even evaluate how strong that belief is – a human can easily predict which state is the most likely future state from Figure 1(c).

The underlying reasoning of the future is more complicated than appearance analysis. The observer needs to understand (i) what happened and what is happening, (ii) what the goal of the agent is, (iii) which object(s) the agent needs to achieve the goal, and (iv) how the agent will perform the task. Based on this rationality, we address the problem of event understanding and human activity prediction from the following two perspectives: (i) a learning algorithm should discover the hierarchical/compositional structure of events, and (ii) an inference algorithm should recover the hierarchical structure given the past observations, and be able to predict the future based on the understanding.

We believe the task of human activity prediction is important for two main reasons. First, the ability to make predictions is key for intelligent systems and robots to perform assistive activities. Second, predicting the future human activities requires deep understanding of human activities. Activity prediction enables the robot to do better task planning. Attempts have been made to address this task in both the computer vision and the robotics community.

In this paper\textsuperscript{1} we aim to design a model that can (i) learn the hierarchical structure of human activities from videos, (ii) online infer the current state of the agent and objects while watching a video, and (iii) predict the next states of the agent and objects. Specifically, the state is defined by the action of the agent, the interacting objects and their affordances, \textit{i.e.} how the objects are being used.

The challenge is three-fold: (i) we need to model the hierarchical structure where the Markov property does not hold. Consider two scenarios: an agent is cleaning the microwave or microwaving food. Whether or not the agent will open the microwave again does not depend on the fact that the agent closed the microwave, but depends on whether or not there is food inside. (ii) Human activities are jointly defined by the human action, the interacting objects, and their affordances. The model needs to capture the spatial-temporal context for event parsing. (iii) We need to predict the human activity from a large future state space.

Inspired by computational linguistics and some recent work in computer vision, we propose a graphical model to represent human activities in a spatial-temporal And-Or graph (ST-AOG), which is composed of a spatial And-Or

\textsuperscript{1}Our full paper\textsuperscript{1} appears at ICCV 2017.
graph (S-AOG) and a temporal And-Or graph (T-AOG). The T-AOG is a stochastic grammar, whose terminal nodes are the root nodes of the spatial graph representing sub-activities. It models the hierarchical structure of human activities and takes the advantage of existing computational linguistic algorithms for symbolic prediction. The S-AOG has child nodes representing a human action, objects, and object affordances. The S-AOG together with T-AOG captures the rich context. For future activity prediction, we first symbolically predict the next sub-activity using the T-AOG, and then predict the human actions and object affordances based on current parsing and sampled future states.

This paper makes three major contributions. i) We propose a spatial-temporal And-Or graph for human activity understanding to incorporate the hierarchical temporal structure and the rich context captured by actions, objects, and affordances. ii) We propose an algorithm for jointly segmenting and parsing the past observations in an online fashion by dynamic programming. iii) We propose a novel algorithm to predict the future human activities. Extensive experiments are conducted to show the effectiveness of our approach by evaluating the classification accuracy of actions and affordances.

2. Representation

We represent the task structure as stochastic context-free grammar using a spatio-temporal And-Or graph (ST-AOG) as shown in Fig. 2. The ST-AOG can be decomposed into two parts: the spatial AOG (S-AOG) and the temporal AOG (T-AOG). The S-AOG is composed of one And-node expanded into a human action, interacting objects and their affordances, representing the human-object interaction for a video segment. The root And-node of an S-AOG is a sub-activity label. The T-AOG is a temporal grammar, in which the root node is the event and the terminal nodes are sub-activities.

Formally, the ST-AOG of an event \( e \in E \) is denoted by \( \mathcal{G}_e = \langle S, \mathcal{V}_{NT} \cup \mathcal{V}_T, \mathcal{R}, \mathcal{P} \rangle \), where \( S \) is root node. \( \mathcal{V}_{NT} \) is the set of non-terminal nodes including the sub-activity labels \( \{S_e\} \). \( \mathcal{V}_T = \{\mathcal{A}_e\}, \{\mathcal{O}_e\}, \{\mathcal{U}_e\} \) is the set of terminal nodes consist of the human action labels \( \{\mathcal{A}_e\} \), the object labels \( \{\mathcal{O}_e\} \), and the affordance labels \( \{\mathcal{U}_e\} \). \( \mathcal{R} \) stands for the production rules, \( \mathcal{P} \) represents the probability model defined on the ST-AOG.

For an event in time \([1, T]\), we extract the skeleton features \( \Gamma_H \), object features \( \Gamma_X \) and the interaction features between the human and the object \( \Gamma_R \) from the video \( I \). We construct a sequence of parse graphs on \( \Gamma = \langle \Gamma_H, \Gamma_X, \Gamma_R \rangle \), which is defined as \( PG = \{p_{t}^{\ell}\}_{t=1,\ldots,T} \). \( PG \) gives us the label \( e \) of the event, and a label sequence \( S = \{s_t\}_{t=1,\ldots,T} \) representing the sub-activity labels of all the frames. We obtain the label sequence \( H = \{h_t\} \), \( O = \{o_t\} \) and \( U = \{u_t\} \) for action, affordance and object labels as well. By merging the consecutive frames with the same sub-activity labels, we obtain the temporal parsing of the video, i.e., \( \mathcal{T} = \{\gamma_k\}_{k=1,\ldots,K} \) where \( \gamma_k = [t^1_k, t^2_k] \) represents a time interval in which the sub-activity remains the same. We use \( \alpha^{\gamma_k}, \alpha^{\gamma_k}_a \), and \( \alpha^{\gamma_k}_o \) to denote the action label, object label and affordance label respectively for video segment \( T^{\gamma_k} \). Both \( a \) and \( o \) are vectors, of which lengths are the number of detected objects.

3. Probabilistic Formulation

In this section, we introduce the probabilistic model defined on the ST-AOG. Given the extracted action, affordance and object features, the posterior probability of a parse graph sequence \( PG \) is defined as:

\[
p(PG|\mathcal{G}_e) \propto p(\Gamma_H|PG)p(\Gamma_X|PG)p(\Gamma_R|PG)p(\mathcal{G}_e|PG)
\]

The first three terms are likelihood terms for actions, objects, and affordances given a parse graph \( PG \). The last term is a prior probability of the parse graph given the grammar \( \mathcal{G} \) of event \( e \).
Figure 3: A simplified example illustrating the parsing and symbolic prediction process. In the first two figures, the red edges and blue edges indicate two different parse graphs for the past observations. The purple edges indicate the overlap of the two possible explanations. The red parse graph is eliminated from the third figure. For the terminal nodes, yellow indicates the current observation and green indicates the next possible state(s).

<table>
<thead>
<tr>
<th>Action</th>
<th>Micro P/R</th>
<th>Micro Recall</th>
<th>Micro F1-score</th>
<th>Macro P/R</th>
<th>Macro Recall</th>
<th>Macro F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>chance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>8.3</td>
<td>8.3</td>
<td>8.3</td>
</tr>
<tr>
<td>LSTM</td>
<td>36.4</td>
<td>30.2</td>
<td>27.7</td>
<td>42.3</td>
<td>36.5</td>
<td>33.4</td>
</tr>
<tr>
<td>VGG-16</td>
<td>46.6</td>
<td>54.0</td>
<td>31.7</td>
<td>40.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KGS [2]</td>
<td>68.2</td>
<td>71.1</td>
<td>62.2</td>
<td>66.4</td>
<td>83.9</td>
<td>75.9</td>
</tr>
<tr>
<td>ATCRF [3]</td>
<td>70.3</td>
<td>74.8</td>
<td>66.2</td>
<td>70.2</td>
<td>85.4</td>
<td>77.0</td>
</tr>
<tr>
<td>ours</td>
<td>76.5</td>
<td>77.0</td>
<td>75.2</td>
<td>76.1</td>
<td>82.4</td>
<td>72.1</td>
</tr>
</tbody>
</table>

Table 1: Detection results on the CAD-120 dataset

4. Learning

The learning of the ST-AOG can be decomposed into two main parts: i) learn the symbolic grammar structure (T-AOG) of each event/task, and ii) learn the parameters $\Theta$ of the ST-AOG, including the branching probabilities of the Or-nodes, the prior distributions of human skeletons and duration of segments. We used a modified version of the ADIOS (automatic distillation of structure) grammar induction algorithm to learn the event grammar. The maximum likelihood estimation (MLE) of the branching probabilities of Or-nodes is simply given by the frequency of each alternative choice [8].

5. Inference

For a single video, we find the parse graph $PG$ for each event $e$ that best explains the extracted features $\Gamma$ by maximizing the posterior probability described in Sec.3:

$$PG = \arg \max_{PG} p(\Gamma_H | A)p(\Gamma_X | O)p(\Gamma_R | U)p(A, O, U | e)$$

This is achieved by i) segment and label the input video by dynamic programming. ii) refine the labels according to the grammar by Gibbs sampling. Then the future activities are made based on Earley parser [1]. We refer the readers to our full paper for more details [4].

6. Experiments and Evaluations

In this section we describe the evaluation of our proposed approach on online parsing and prediction. We perform our experiments on CAD-120 dataset [2]. It has 120 RGB-D videos of four different subjects performing 10 activities, each of which is a sequence of sub-activities involving 10 actions (e.g. reaching, opening), and 12 object affordance (e.g. reachable, openable) in total. We compare our method with recently proposed methods [3 2] and several other baselines.

6.1. Parsing Results

We parsed the videos frame by frame in an online fashion and evaluated the detection results for the current frame.
The model is trained on three subjects and tested on a new subject. Results are obtained by four-fold cross validation by averaging across the folds. We trained the action and affordance detectors using a simple two-layer fully connected neural network based on features similar to [3]. We fine-tuned Faster R-CNN [5] for object detection. We compared our detection results with the following methods: 1) Chance. The labels are chosen randomly. 2) SVM: An SVM trained on our features. 3) LSTM: A two-layer LSTM trained on our features. 4) VGG-16 [6]: Using the image as input, we fine-tuned a VGG-16 network on the action labels. Since the object affordances are evaluated on each object instead of an image (an image can have multiple objects thus can have multiple affordance labels), we only evaluate the performance of action detection. 5) KGS [2]: A Markov random field model where the nodes represent objects and sub-activities, and the edges represent the spatial-temporal relationships. 6) ATCRF [3]: An anticipatory temporal conditional random field that models the spatial-temporal relations through object affordances.

Figure 4 shows the confusion matrix for classifying actions and affordances, and we report the overall micro accuracy, macro precision and macro recall of the detected actions and affordances in Table 1. Our approach outperforms the other methods on action detection, and achieves a comparable performance with ATCRF [3] on affordance detection.

In the experiments, we found that the algorithm is generally capable of improving the low-level detections using joint high-level reasoning. For example, one “stacking objects” video has an input action detection accuracy of 50.9% and affordance detection accuracy of 84.5%. After joint reasoning, the output action detection accuracy raised to 86.7% and affordance detection accuracy raised to 87.3%.

6.2. Prediction Results

We report the frame-wise accuracy of prediction on actions and affordances over 3 seconds in the future (using frame rate of 14Hz as reported in [3]). Table 2 shows the comparisons between our approach and other methods. We achieved a better performance for all predictions even though the detection result is not the best.

Qualitative results Based on the predicted affordance labels, we can predict which object human is going to interact with. Figure 5 shows the predicted right hand trajectory heat maps within the next one second.

References