Learning Articulated Object Models from Language and Vision

Andrea F. Daniele
TTI-Chicago
Chicago, IL, USA
afdaniele@ttic.edu

Thomas M. Howard
University of Rochester
Rochester, NY, USA
thoward@ece.rochester.edu

Matthew R. Walter
TTI-Chicago
Chicago, IL, USA
mwalter@ttic.edu

Abstract

In order for robots to operate effectively in homes and workplaces, they must be able to manipulate the articulated objects common within environments built for and by humans. Kinematic models provide a concise representation of these objects that enable deliberate, generalizable manipulation policies. However, existing approaches to learning these models rely upon visual observations of an object's motion, and are subject to the effects of occlusions and feature sparsity. Natural language descriptions provide a flexible, efficient means by which humans can provide complementary information in a weakly supervised manner. We present a multimodal learning framework that incorporates both vision and language information acquired in situ to estimate the structure and parameters that define kinematic models of articulated objects. We model linguistic information using a probabilistic language model that grounds natural language descriptions to their referent kinematic motion. By exploiting the complementary nature of vision and language, our method infers correct kinematic models for various multiple-part objects on which the previous state-of-the-art, visual-only system fails. We evaluate our multimodal learning framework on a dataset comprised of a variety of household objects, and demonstrate 23% improvement in accuracy over the vision-only baseline.

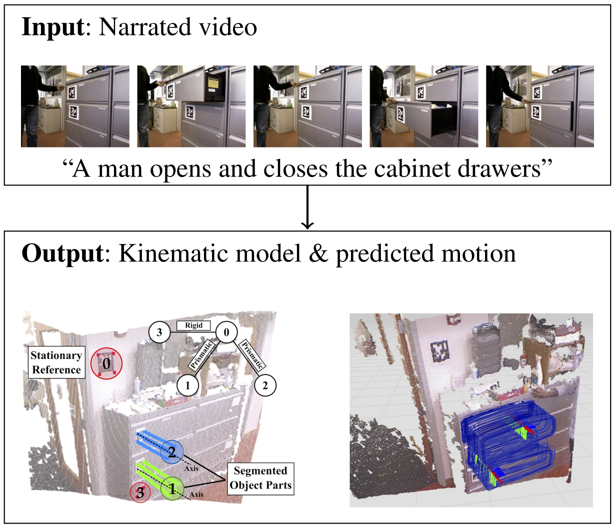
Input: Narrated video
"A man opens and closes the cabinet drawers"
Output: Kinematic model & predicted motion
Stationary Reference
0
3
Rigid
0
Prismatic
Prismatic
1
2
2
Axis
1
Axis
3
Segmented Object Parts

Figure 1: Our framework learns the kinematic model that governs the motion of articulated objects (lower-left) from narrated RGB-D videos. It then uses this learned model to predict the motion of an object's parts (lower-right).

Introduction

As robots move off factory floors and into our homes and workplaces, they face the challenge of interacting with the articulated objects frequently found in environments built by and for humans (e.g., drawers, ovens, refrigerators, and faucets). Typically, this interaction is predefined in the form of a manipulation policy that must be (manually) specified for each object that the robot is expected to interact with. Such an approach may be reasonable for robots that interact with a small number of objects, but human environments contain a large number of diverse objects. In an effort to improve efficiency and generalizability, recent work employs visual demonstrations to learn representations that describe the motion of these parts in the form of kinematic models that express the rotational, prismatic, and rigid relationships between object parts [6, 16, 20, 35, 40]. These structured object-relative models, which constrain the object's motion manifold, are suitable for trajectory controllers [17, 40], provide a common representation amenable to transfer between objects [41], and allow for manipulation policies that are more efficient and deliberate than reactive policies. However, such visual cues may be too time-consuming to provide or may not be readily available, such as when a user is remotely commanding a robot over a bandwidth-limited channel (e.g., for disaster relief). Further, reliance solely on vision makes these methods sensitive to common errors in object segmentation and tracking that occur as a result of clutter, occlusions, and a lack of visual features. Consequently, most existing systems require scenes to be free of distractors and that object parts be labeled with fiducial markers.

Natural language utterances offer a flexible, bandwidth-efficient medium that humans can readily use to convey knowledge of an object's operation [41]. When paired with visual observations, such as in the case of instructional videos [27, 39, 49], free-form descriptions of an articulated motion also provide a source of information that is complementary to visual input. Thus, these descriptions can be used to overcome some of the limitations of using visual-only observations, e.g., by providing cues regarding the number of parts that comprise the object or the motion type between a pair of parts. However, fusing visual and linguistic observations is challenging. For one, language and vision provide disparate observations of motion and exhibit different statistical properties. Secondly, the two are often prone to uncertainty. RGB-D observations are subject to occlusions (e.g., as the human interacts with the object) and the objects often lack texture (e.g., the drawers in Fig. 1), which makes feature detection challenging and feature correspondences subject to noise. Meanwhile, free-form descriptions exhibit variability and are prone to errors (e.g., confusing "left" and "right"). Further, language also tends to be ambiguous with respect to the corresponding referents (i.e., object parts and their motion) in the scene. For example "open" can imply rotational and prismatic motion and inferring the correct grounding requires reasoning over the full description (e.g., "open the door" vs. "open the cabinet drawers").

In order to overcome these challenges, we present a multimodal learning framework that estimates the kinematic structure and parameters of complex multi-part objects using both vision and language input. We address the challenges associated with language understanding through a probabilistic language model that captures the compositional and hierarchical structure of natural language descriptions. Additionally, our method maintains a distribution over a sparse, structured model of an object's kinematics, which provides a common representation with which to fuse disparate linguistic and visual observations.

Our effort is inspired by the recent attention that has been paid to the joint use of vision and language as complementary signals for multiview learning in robotics [13, 23, 36, 41, 44, 50] and scene understanding [1, 11, 19, 25, 37, 43]. We leverage the joint advantages of these two modalities in order to estimate the structure and parameters that define kinematic models of complex, multi-part objects such as doors, desks, chairs, and appliances from narrated examples such as those conveyed in instructional videos or through demonstrations [3] in the form of a "guided tour of manipulation" (Fig. 1), which provides an efficient and flexible means for humans to share information with robots.

Our multimodal learning framework first extracts noisy observations of the object parts and their motion separately from the vision- and language-based observations. It then fuses these observations to learn a probabilistic model over the kinematic structure and model parameters that best explain the motion observed in the vision and language streams. Integral to this process is an appropriate means of representing the ambiguous nature of observations gleaned from natural language descriptions. We treat language understanding as a symbol grounding problem and employ a probabilistic language model [15] that captures the uncertainty in the mapping between words in the description and their corresponding referents in the scene, namely the object parts and their relative motion. We fuse these language-based observations with those extracted from vision to estimate a joint distribution over the structure and parameters that define the kinematics of each object.

The contributions of this work include a multimodal approach to learning kinematic models from vision and language signals and the integration of a probabilistic language model that grounds natural language descriptions into a structured representation of an object's articulation manifold. By jointly reasoning over vision and language cues, our framework is able to formulate a complete object model without the need for an expressed environment model. Our method requires no prior knowledge about the objects and operates in situ, without the need for environment preparation (i.e., fiducials). Evaluations on a dataset of video-text pairs demonstrate improvements over the previous state-of-the-art, which only uses visual information.

Related Work

Recent work considers the problem of learning articulated models based upon visual observations of demonstrated motion. Several methods formulate this problem as bundle adjustment, using structure-from-motion methods to first segment an articulated object into its compositional parts and to then estimate the parameters of the rotational and prismatic degrees-of-freedom that describe inter-part motion [16, 48]. These methods are prone to erroneous estimates of the pose of the object's parts and of the inter-part models as a result of outliers in visual feature matching. Alternatively, Katz, Orthey, and Brock [21] propose an active learning framework that allows a robot to interact with articulated objects to induce motion. This method operates in a deterministic manner, first assuming that each part-to-part motion is prismatic. Only when the residual error exceeds a threshold does it consider the alternative rotational model. Further, they estimate the models based upon interactive observations acquired in a structured environment free of clutter, with the object occupying a significant portion of the RGB-D sensor's field-of-view. Katz et al. [20] improve upon the complexity of this method while preserving the accuracy of the inferred models. This method is prone to over-fitting to the observed motion and may result in overly complex models to match the observations. Hausman et al. [12] similarly enable a robot to interact with the object and describe a probabilistic model that integrates observations of fiducials with manipulator feedback. Meanwhile, Sturm, Stachniss, and Burgard [40] propose a probabilistic approach that simultaneously reasons over the likelihood of observations while accounting for the learned model complexity. Their method requires that the number of parts that compose the object be known in advance and that fiducials be placed on each part to enable the visual observation of motion. More recently, Pillai, Walter, and Teller [35] propose an extension to this work that uses novel vision-based motion segmentation and tracking that enables model learning in situ, without prior knowledge of the number of parts or the need for fiducial markers. Our approach builds upon this method with the addition of natural language descriptions of motion as an additional observation mode in a multimodal learning framework. Meanwhile, Schmidt, Newcombe, and Fox [38] use an articulated variation of the signed distance function to identify the model that best fits observed depth data.

Related, there has been renewed attention to enabling robots to interpret natural language instructions that command navigation [4, 7, 22, 29, 30] and manipulation [15, 31, 34, 42] through symbol grounding and semantic parsing methods. While most existing grounded language acquisition methods abstract away perception by assuming a known symbolic world model, other work jointly reasons over language and sensing [8, 10, 14, 28] for instruction following. Meanwhile, multimodal learning methods have been proposed that use language and vision to formulate spatial-semantic maps of a robot's environment [13, 36, 44, 50] and to learn object manipulation policies [41]. Particularly relevant to our work, Sung, Jin, and Saxena [41] learn a neural embedding of text, vision, and motion trajectories to transfer manipulation plans between similarly operating objects. Kollar, Krishnamurthy, and Strimel [23] extend their framework that jointly learns a semantic parsing of language and vision [25] to enable robots to learn object and spatial relation classifiers from textual descriptions paired with images. We similarly use language and vision in a joint learning framework, but for the challenging task of learning object articulation in terms of kinematic motion models. Beyond robotics, there is a long history of work that exploits the complementary nature of vision and language in the context of multiview learning, dating back to the seminal SHRDLU program [46]. This includes work for such tasks as image and video caption synthesis [19, 33, 43, 47], large-vocabulary object retrieval [11], visual coreference resolution [24, 37], and visual question-answering [2]. Particularly related with our work are methods that use instructional videos paired with language (text or speech) for weakly supervised learning [27, 49], extracting procedural knowledge [39], and identifying manipulating actions [1, 41].

Multimodal Learning Framework

Given an RGB-D video paired with the corresponding natural language description (alternatively, an instruction or caption) of an articulated object's motion, our goal is to infer the structure and parameters of the object's kinematic model. Adopting the formulation proposed by Sturm, Stachniss, and Burgard [40], we represent this model as a graph, where each vertex denotes a different part of the object (or the stationary background) and edges denote the existence of constrained motion (e.g., a linkage) between two parts (Fig. 1). More formally, we estimate a *kinematic graph* $G = (V_G, E_G)$ that consists of vertices V_G for each object part and edges $E_G \subset V_G \times V_G$ between parts whose relative motion is kinematically constrained. Associated with each edge $(ij) \in E_G$ is its kinematic type $M_{ij} \in \{\text{rotational}, \text{prismatic}, \text{rigid}\}$ as well as the corresponding parameters θ_{ij}, such as the axis of rotation and the range of motion (see Fig. 2, lower-right). We take as input vision D_v and language D_l observations of the type and parameters of the edges in the graph. Our method then uses this vision-language observation pair $D_z = \{D_v, D_l\}$ to infer the maximum a posteriori kinematic structure and model parameters that constitute the kinematic graph:

$$\hat{G} = \arg\max_G \, p(G|D_z) \tag{1a}$$

$$= \arg\max_G \, p(\{M_{ij}, \theta_{ij} | (ij) \in E_G\} | D_z) \tag{1b}$$

$$= \arg\max_G \prod_{(ij)\in E_G} p(M_{ij}, \theta_{ij} | D_z) \tag{1c}$$

Due to the complexity of joint inference, we adopt the procedure described by Sturm, Stachniss, and Burgard [40] and use a two-step inference procedure that alternates between model parameter fitting and model structure selection steps (Fig. 2). In the first step, we assume a particular kinematic model type between each object i and j (e.g., prismatic), and then estimate the kinematic parameters based on the vision data (relative transformation between the two objects) and the assumed model type M_{ij}. We make one assumption for each possible model type for each object pair.

In the model selection step, we then use the natural language description to infer the kinematic graph structure that best expresses the observation. While our previous work [35] provides visual observations of motion without the need for fiducials, it relies upon feature tracking and segmentation that can fail when the object parts lack texture (e.g., metal door handles) or when the scene is cluttered. Our system incorporates language as an additional, complementary observation of the motion, in order to improve the robustness and accuracy of model selection.

Vision-guided Model Fitting

We parse a given RGB-D video of the objects motion (either performed by a human or the robot via teleoperation) to arrive at a visual observation of the trajectory of each object part [35]. The method (Fig. 2, "Construct Trajectories") first identifies a set of 3D feature trajectories $\{(f_1^1, f_1^2, \ldots, f_1^t), \ldots (f_n^1, f_n^2, \ldots, f_n^t)\}$ that correspond to different elements in the scene, including the object parts, background, and clutter. Importantly, both the number of elements and the assignment of points in the RGB-D video to these elements are assumed to be unknown a priori. Further, many of the objects that we encounter lack the amount of texture typically required of SIFT [26] and KLT [5] features. Consequently, we utilize dense trajectories [45] through a strategy that involves dense sampling (via the Shi-Tomasi criterion) for feature extraction followed by dense optical flow for propagation. We prune trajectories after a fixed length and subsequently sample new features to reduce drift.

Having extracted a set of features trajectories, the next step is then to group features that correspond to the same scene element via motion segmentation (Fig. 2, "Motion Segmentation"). For this purpose, we evaluate the relative displacement between pairs of feature trajectories along with the angle between their normals. We model the relative displacement and angle as Gaussian in order to account for measurement noise. We then employ density-based clustering [9] to identify rigidly associated feature trajectories. These clusters $\{C_1, C_2, \ldots, C_m\}$ denote the parsing of the scene into its requisite elements, namely the inferred object parts and background.

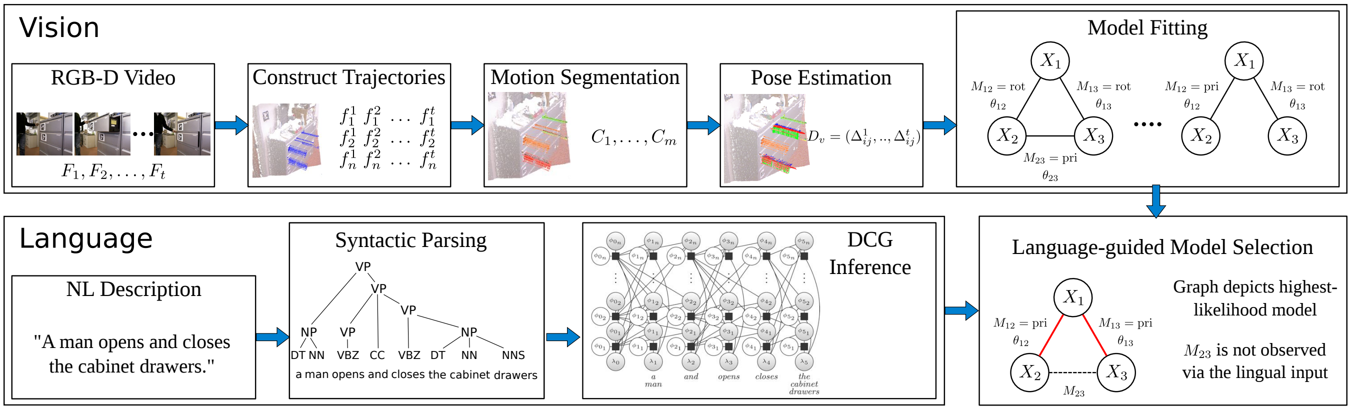

Figure 2: Our multimodal articulation learning framework first identifies clusters of visual features that correspond to individual object parts. It then uses these feature trajectories to estimate the model parameters, assuming an initial estimate of the kinematic type associated with each edge in the graph. The method grounds natural language descriptions of the motion to their corresponding referents in the kinematic model and parameters through a probabilistic language model, visualized as a factor graph. The vision and language observations are then fused to learn a distribution over the object's kinematic model.

Next, we estimate the 6-DOF pose x_i^t of each cluster at each point in time according to the set of features Z_i^t assigned to each cluster C_i at time t (Fig. 2, "Pose Estimation"). We treat this as a pose graph estimation problem, whereby we consider the relative transformation $\Delta_i^{t-1,t}$ between successive time steps for each cluster based on the known correspondences between features Z_i^{t-1} and Z_i^t.We optimize the pose graph using iSAM [18], which models the relative transformations as observations (constraints) in a factor graph with nodes that denote cluster poses.

The resulting 6-DOF pose trajectories constitute the visual observation of the motion D_v. Our framework uses these trajectories to estimate the parameters of a candidate kinematic model during the model fitting step. Specifically, we find the kinematic parameters that best explain the visual data given the assumed model

$$\hat{\theta}_{ij} = \underset{\theta_{ij}}{\arg\max}\, p(D_v | \hat{M}_{ij}, \theta_{ij}), \tag{2}$$

where $D_v = (\Delta_{ij}^1, ..., \Delta_{ij}^t), \forall (ij) \in E_G$ is the sequence of observed relative transformations between the poses of two object parts i and j, and $\hat{M}_{ij}$ is the current estimate of their model type. We perform this optimization over the joint kinematic structure defined by the edges in the graph [40].

Language-guided Model Selection

Methods that solely rely on visual input are sensitive to the effects of scene clutter and the lack of texture, which can result in erroneous estimates for the structure and parameters of the kinematic model [35]. An alternative is to exploit audial information, provided in the form of utterances provided by the operator, to help guide the process for inferring the relationships between objects in the environment. Specifically, we consider a natural language description D_l that describes the motion observed in the video. Given this description, we infer the maximum a posteriori set of affordances or relationships between pairwise objects in the utterance. Note that we do not assume that valid captions provide an unambiguous description of all affordances, but rather consider a distribution over the observation, which provides robustness to noisy, incomplete, or incorrect descriptions.

Following the notation from Paul et al. [34], we formulate this problem as one where we must infer a distribution of symbols (Γ) representing objects ($\Gamma^{\mathcal{O}}$), relationships ($\Gamma^{\mathcal{R}}$), and affordances ($\Gamma^{\mathcal{A}}$) in the absence of an environment model for each utterance. Object groundings are defined by an object type o_i from a space of object types $\mathcal{O}$, relationship groundings are defined by an relationship type r_k from a space of relationship types $\mathcal{R}$, and affordance groundings are defined by a pair of object types o_i and o_j, and relationship type r_k:

$$\Gamma^{\mathcal{O}} = \{\gamma_{o_i}, o_i \in \mathcal{O}\} \tag{3a}$$

$$\Gamma^{\mathcal{R}} = \{\gamma_{r_k}, r_k \in \mathcal{R}\} \tag{3b}$$

$$\Gamma^{\mathcal{A}} = \{\gamma_{o_i,o_j,r_k}, o_i, o_j \in \mathcal{O}, r_k \in \mathcal{R}\} \tag{3c}$$

Examples of object types include "chair", "desk", and "door" which represent semantic classes of random variables inferred by visual perception. Examples of relationship types include "prismatic" and "revolute" that represent translational and rotational motion. The set of all groundings is defined as the union of these symbols:

$$\Gamma = \{\Gamma^{\mathcal{O}} \cup \Gamma^{\mathcal{D}} \cup \Gamma^{\mathcal{A}}\} \tag{4}$$

Extracting the most probable set of groundings from language is challenging due to the diversity inherent in free-form language and the complex relationships between the articulation of different objects. For example, the verb "open" can be used to describe a person's interaction with both a drawer and door, but the motion described in the former case is prismatic with a cabinet, while it is rotational with a wall in the latter. We address these challenges by adapting the Distributed Correspondence Graph (DCG) [15] to the problem of affordance inference which formulates a probabilistic graphical model according to the parse structure of the sentence that is searched for the most likely binary correspondence variable $\phi_{i,j} \in \{\text{TRUE}, \text{FALSE}\}$ between linguistic elements in the command λ_i, groundings $\gamma_{i,j} \in \Gamma$, and expressed groundings of child phrases $\Gamma_{c_i} \in \Gamma$. The DCG encodes the factors $f_{i,j}$ in the graph using log-linear models whose weights are learned from a corpus of annotated examples. We then perform inference over this model in a space of correspondence variables to arrive at a distribution over the kinematic model structure and parameters. Note that since inference is conducted without an expressed environment model, the symbols express inferred relationships between semantic classes of pairwise objects are interleaved with the model constructed from visual perception to form a probabilistic model of the environment. An example of the structure of the DCG for the utterance "a man opens and closes the cabinet drawers" is illustrated in Figure 3.

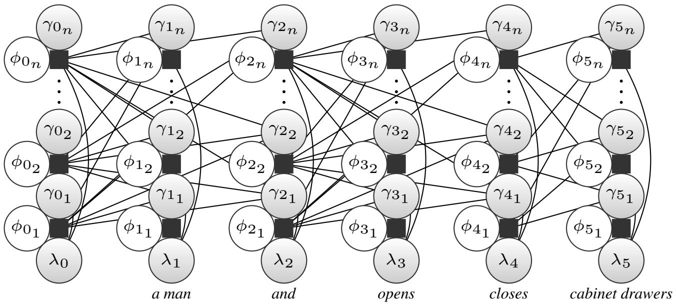

Figure 3: The DCG for the utterance "a man opens and closes the cabinet drawers" constructed from the parse tree illustrated in Figure 2. This model enumerates all possible groundings for each phrase and performs inference by searching over unknown correspondences. The expressed groundings (groundings for factors with TRUE-valued correspondence variables) of factors connected to λ_0 are used as language-based observations that are fused with the visual observation. The method infers a "prismatic" relationship between objects of semantic classes "cabinet" and "door".

Combining Vision and Language Observations

The final step in our framework selects the kinematic graph structure $\hat{\mathcal{M}} = \{\hat{M}_{ij}, \forall (ij) \in E_G\}$ that best explains the vision and language observations $D_z = \{D_v, D_l\}$ from the space of all possible kinematic graphs. We do so by maximizing the conditional posterior over the model type associated with each edge in the graph $(ij) \in E_G$:

$$\hat{M}_{ij} = \arg\max_{M_{ij}} p(M_{ij} | D_z) \tag{5a}$$

$$= \arg\max_{M_{ij}} \int p(M_{ij}, \theta_{ij} | D_z) d\theta_{ij} \tag{5b}$$

Evaluating this likelihood is computationally prohibitive, so we use the Bayesian Information Criterion (BIC) score as an approximation

$$BIC(M_{ij}) = -2 \log p(D_z | M_{ij}, \hat{\theta}_{ij}) + k \log n, \tag{6}$$

where $\hat{\theta}_{ij}$ is the maximum likelihood parameter estimate (Eqn. 2), k is the number of parameters of the current model and n is the number of vision and language observations. We choose the model with the lowest BIC score

$$\hat{M}_{ij} = \arg\min_{M_{ij}} BIC(M_{ij}) \tag{7}$$

as that which specifies the kinematics of the object.

While our previous method [35] only considers visual measurements, our new framework performs this optimization over the joint space of vision and language observations. Consequently, the BIC score becomes

$$\begin{aligned} BIC(M_{ij}) = & -2\Big(\log p(D_v | M_{ij}, \hat{\theta}_{ij}) + \log p(D_l | M_{ij}, \hat{\theta}_{ij})\Big) \\ & + k \log n \end{aligned}$$

where we have made the assumption that the language and vision observations are conditionally independent given the model and parameter estimates. We formulate the conditional likelihood of the linguistic observation according to the grounding likelihood $P(\Phi = \text{TRUE} | \gamma_1, \ldots, \gamma_n, \Lambda)$ from the DCG language model. The grounding variables γ_i denote affordances that express different kinematic structures that encode the articulation of the object, namely the relationship between its individual parts (Eqn. 3). For each candidate model, we use the likelihood of the corresponding groundings under the learned DCG language model to compute the BIC score for the corresponding affordance. We then estimate the overall kinematic structure by solving for the minimum spanning tree of the graph, where we define the cost of each edge as $\text{cost}_{ij} = -\log p(M_{ij}, \theta_{ij} | D_z)$. Such a spanning tree constitutes the kinematic graph that best describes the vision and language observations.

Results

We evaluate our framework using a dataset of 78 RGB-D videos in which a user manipulates a variety of common household and office objects (e.g., a microwave, refrigerator, and filing cabinet). Each video is accompanied with 5 textual descriptions provided by different human subjects using a web-based crowd-sourcing platform. We split the dataset into separate training and test sets consisting of 22 and 56 videos, respectively. AprilTags [32] were placed on each of the object parts in the test set to determine ground-truth motion. We train our language grounding model on a corpus of 50 video descriptions corresponding to 28 unique symbols composed of different object and/or relation types.

Of the 56 test videos, 25 involve single-part objects and 31 involve multi-part objects. The single-part object videos are used to demonstrate that the addition of language observations can only improve the accuracy of the learned kinematic models. The extent of these improvements on single-part objects is limited by the relative ease of inference of single degree-of-freedom motion. In the case of multi-part objects, the larger space of candidate kinematic graphs makes vision-only inference challenging, as feature tracking errors may result in erroneous estimates of the graph structure.

Evaluation Metrics and Baselines

We estimate the ground-truth kinematic models by performing MAP inference based upon the motion trajectories observed using AprilTags. We denote the resulting kinematic graph as G^*. The kinematic type and parameters for each object part pair are denoted as M^*_{ij} and θ^*_{ij}, respectively. Let $\hat{G}$, $\hat{M}_{ij}$, $\hat{\theta}_{ij}$ be the estimated kinematic graph, kinematic type, and parameters for each object pair from the RGB-D video.

The first metric that we consider evaluates whether the vision component estimates the correct number of parts. We determine the ground-truth number of parts as the number of AprilTags observed in each video, which we denote as N^*. We indicate the number of parts (motion clusters) identified by the visual pipeline as N_v. We report the average success rate when using only visual observations as $S_v = \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}(N_v^k = N^{k*})$, where K is the number of videos for each object type.

Next, we consider two metrics that assess the ability of each method to estimate a graph with the same kinematic model as the ground truth G^*. The first metric requires that the two graphs have the same structure, i.e., $\hat{M}_{ij} = M^*_{ij}, \forall (ij) \in E_{\hat{G}} = E_{G^*}$. This equivalence requires that vision-only inference yields the correct number of object parts and that the model selection framework selects the correct kinematic edge type for each pair of object parts. We report this "hard" success rate S_h in terms of the fraction of demonstrations for which the model estimate agrees with ground truth. Note that this is bounded from above by fraction for which the vision component estimates the correct number of parts. The second "soft" success rate (denoted by S_s) employs a relaxed requirement whereby we only consider the inter-part relationships identified from vision, i.e., $\hat{M}_{ij} = M^*_{ij}, \forall (ij) \in E_{\hat{G}} \subset E_{G^*}$. In this way, we consider scenarios for which the visual system detects fewer parts than are in the ground-truth model. In our experiments, we found that $\hat{G}$ is a sub-graph of G^*, so we only require that the model type of the edges in this sub-graph agree between both graphs. The metric reports the fraction of total demonstrations for which the estimated kinematic graph is a correct sub-graph of the ground-truth kinematic graph.

Once we have the same kinematic models for both $\hat{G}$ and G^*, we can compare the kinematic parameters $\hat{\theta}_{ij}$ to the ground-truth values θ^*_{ij} for each inter-part model $\hat{M}_{ij}$. Note that for the soft metric, we only compare kinematic parameters for edges in the sub-graph, i.e., $\forall (ij) \in E_{\hat{G}} \subset E_{G^*}$. We define the parameter estimation error for a particular part pair as the angle between the two kinematic parameter axes,

$$\mathrm{e}_{ij} = \arccos \frac{\hat{\theta}_{ij} \cdot \theta^*_{ij}}{\|\hat{\theta}_{ij}\|\|\theta^*_{ij}\|}, \tag{8}$$

where we use the directional and rotational axes for prismatic and rotational degrees-of-freedom, respectively. We measure the overall parameter estimation error e_{param} for an object as the average parameter estimation error over each edge in the object's kinematic graph. We report this error further averaged over the number of demonstrations.

Results and Analysis

Table 1 summarizes the performance of our multimodal learning method using our embedding-based language model with hard alignment, comparing against the performance of the vision-only baseline [35]. The table indicates the number of demonstrations (K), the ground-truth number of parts for each object (N^*), a list of the number of parts identified using visual trajectory clustering for each demonstration (N_v), and the fraction of videos for which the correct number of parts was identified (S_v). We then present the hard (S_h) and soft (S_s) model selection rates for our method as well as for the baseline. Our method bests the vision-only baseline in estimating the full kinematic graph for five of the eight objects, matching its performance on the remaining three objects. Specifically, our framework yields accurate estimates of the full kinematic graphs for thirteen more demonstrations than the vision-only baseline, nine more for single-part objects and four more for multi-part objects, corresponding to a 23% absolute improvement. Similarly, we are able to estimate a valid sub-graph of the ground-truth kinematic graph for eighteen more demonstrations than the vision-only baseline (eleven for single-part and seven for multi-part objects), corresponding to a 19% absolute improvement. One notable object on which both methods have difficulty is the bicycle for which the trajectory clustering method was unable to identify the presence of the third part (the wheel) due to the sparsity of visual features on the wheels. Consequently, neither method estimated the full kinematic graph for any video. Similarly, clustering failed to identify the three parts that comprise the monitor in all the videos, however our framework was able to exploit language to estimate an accurate sub-graph for one more video.

We then evaluate the accuracy of the parameters estimated by our method by reporting the parameter estimation error for each object, averaged over the set of videos. Note that it is difficult to compare against the error of the vision-only baseline since it does not yield accurate kinematic graphs for several of the videos. When the kinematic graph estimates agree, however, the parameter estimation errors are identical for the two methods, since they both estimate the parameters from the visual data (Eqn. 2).

Conclusion

We described a method that uses a joint combination of vision- and language-based observations to learn accurate probabilistic models that define the structure and parameters of articulated objects. Our framework treats descriptions of a demonstrated motion as a complementary observation of the structure of kinematic linkages. We evaluate our framework on a series of RGB-D video-description pairs involving the manipulation of common household objects. The results demonstrate that exploiting language as a form of weak supervision improves the accuracy of the inferred model structure and parameters. Future work includes incorporating semantic segmentation to assign perceived labels to inferred clusters, using the description to mitigate noise in the visual recognition.

Table 1: Overall performance of our framework on video-description pairs.

						Vision-Only		Our Framework	
	Object	K	N^*	N_v	S_v	S_h	S_s	S_h	S_s
Single-Part	Door	9	1	1(9)	9/9	5/9	5/9	**9/9**	**9/9**
	Chair	5	1	1(4), 3	4/5	1/5	2/5	**4/5**	**5/5**
	Refrigerator	5	1	1(5)	5/5	5/5	5/5	5/5	5/5
	Microwave	4	1	1(3), 2	3/4	3/4	4/4	3/4	4/4
	Drawer	2	1	1(2)	2/2	0/2	0/2	**2/2**	**2/2**
Multi-Part	Chair	4	2	1(2), 2(2)	2/4	1/4	5/8	**2/4**	**6/8**
	Monitor	7	2	1(7)	0/7	0/7	6/14	0/7	**7/14**
	Bicycle	7	3	1(1), 2(4), 3(2)	2/7	0/7	13/21	0/7	13/21
	Drawer	11	2	1(6), 2(4), 3(1)	4/11	3/13	10/24	**6/13**	**17/24**
	Door	2	2	2(2)	2/2	0/2	2/4	**2/2**	**4/4**

References

[1] Alayrac, J.-B.; Sivic, J.; Laptev, I.; and Lacoste-Julien, S. 2017. Joint discovery of object states and manipulation actions. *arXiv preprint arXiv:1702.02738*.

[2] Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual question answering. *arXiv preprint arXiv:1505.00468*.

[3] Argall, B. D.; Chernova, S.; Veloso, M.; and Browning, B. 2009. A survey of robot learning from demonstration. *Robotics and Autonomous Systems* 57(5):469–483.

[4] Artzi, Y., and Zettlemoyer, L. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Trans. Assoc. for Computational Linguistics* 1:49–62.

[5] Bouguet, J.-Y. 2001. Pyramidal implementation of the affine Lucas-Kanade feature tracker description of the algorithm. *Intel Corporation*.

[6] Byravan, A., and Fox, D. 2017. SE3-Nets: Learning rigid body motion using deep neural networks. In *Proc. ICRA*.

[7] Chen, D. L., and Mooney, R. J. 2011. Learning to interpret natural language navigation instructions from observations. In *Proc. AAAI*.

[8] Duvallet, F.; Walter, M. R.; Howard, T.; Hemachandra, S.; Oh, J.; Teller, S.; Roy, N.; and Stentz, A. 2014. Inferring maps and behaviors from natural language instructions. In *Proc. ISER*.

[9] Ester, M.; Kriegel, H.-P.; Sander, J.; and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. KDD*, 226–231.

[10] Guadarrama, S.; Riano, L.; Golland, D.; Gohring, D.; Jia, Y.; Klein, D.; Abbeel, P.; ; and Darrell, T. 2013. Grounding spatial relations for human-robot interaction. In *Proc. IROS*, 1640–1647.

[11] Guadarrama, S.; Rodner, E.; Saenko, K.; Zhang, N.; Farrell, R.; Donahue, J.; and Darrell, T. 2014. Open-vocabulary object retrieval. In *Proc. RSS*.

[12] Hausman, K.; Niekum, S.; Ostenoski, S.; and Sukhatme, G. S. 2015. Active articulation model estimation through interactive perception. In *Proc. ICRA*, 3305–3312.

[13] Hemachandra, S.; Walter, M. R.; Tellex, S.; and Teller, S. 2014. Learning spatially-semantic representations from natural language descriptions and scene classifications. In *Proc. ICRA*.

[14] Hemachandra, S.; Duvallet, F.; Howard, T. M.; Roy, N.; Stentz, A.; and Walter, M. R. 2015. Learning models for following natural language directions in unknown environments. In *Proc. ICRA*.

[15] Howard, T. M.; Tellex, S.; and Roy, N. 2014. A natural language planner interface for mobile manipulators. In *Proc. ICRA*, 6652–6659.

[16] Huang, X.; Walker, I.; and Birchfield, S. 2012. Occlusion-aware reconstruction and manipulation of 3D articulated objects. In *Proc. ICRA*, 1365–1371.

[17] Jain, A., and Kemp, C. C. 2010. Pulling open doors and drawers: Coordinating an omnidirectional base and a compliant arm with equilibrium point control. In *Proc. ICRA*.

[18] Kaess, M.; Ranganathan, A.; and Dellaert, F. 2008. iSAM: Incremental smoothing and mapping. *Trans. on Robotics* 24(6):1365–1378.

[19] Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proc. CVPR*.

[20] Katz, D.; Kazemi, M.; Andrew Bagnell, J.; and Stentz, A. 2013. Interactive segmentation, tracking, and kinematic modeling of unknown 3D articulated objects. In *Proc. ICRA*, 5003–5010.

[21] Katz, D.; Orthey, A.; and Brock, O. 2010. Interactive perception of articulated objects. In *Proc. ISER*.

[22] Kollar, T.; Tellex, S.; Roy, D.; and Roy, N. 2010. Toward understanding natural language directions. In *Proc. HRI*, 259–266.

[23] Kollar, T.; Krishnamurthy, J.; and Strimel, G. 2013. Toward interactive grounded language acquisition. In *Proc. RSS*.

[24] Kong, C.; Lin, D.; Bansal, M.; Urtasun, R.; and Fidler, S. 2014. What are you talking about? text-to-image coreference. In *Proc. CVPR*.

[25] Krishnamurthy, J., and Kollar, T. 2013. Jointly learning to parse and perceive: Connecting natural language to the physical world. *Transactions of the Association for Computational Linguistics* 193–206.

[26] Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *Int'l J. on Computer Vision* 60(2):91–110.

[27] Malmaud, J.; Huang, J.; Rathod, V.; Johnston, N.; Rabinovich, A.; and Murphy, K. 2015. What's cookin'? Interpreting cooking videos using text, speech and vision. *arXiv preprint arXiv:1503.01558*.

[28] Matuszek, C.; Herbst, E.; Zettlemoyer, L.; and Fox, D. 2012. Learning to parse natural language commands to a robot control system. In *Proc. ISER*, 403–415.

[29] Matuszek, C.; Fox, D.; and Koscher, K. 2010. Following directions using statistical machine translation. In *Proc. HRI*.

[30] Mei, H.; Bansal, M.; and Walter, M. R. 2016. Listen, attend, and walk: Neural mapping of navigational instructions to action sequences. In *Proc. AAAI*.

[31] Misra, D. K.; Sung, J.; Lee, K.; and Saxena, A. 2016. Tell me Dave: Context-sensitive grounding of natural language to manipulation instructions. *Int'l J. of Robotics Research* 35:281–300.

[32] Olson, E. 2011. AprilTag: A robust and flexible visual fiducial system. In *Proc. ICRA*.

[33] Ordonez, V.; Kulkarni, G.; and Berg, T. L. 2011. Im2Text: Describing images using 1 million captioned photographs. In *(NIPS)*.

[34] Paul, R.; Arkin, J.; Roy, N.; and Howard, T. M. 2016. Efficient grounding of abstract spatial concepts for natural language interaction with robot manipulators. In *Proc. RSS*.

[35] Pillai, S.; Walter, M. R.; and Teller, S. 2014. Learning articulated motions from visual demonstration. In *Proc. RSS*.

[36] Pronobis, A., and Jensfelt, P. 2012. Large-scale semantic mapping and reasoning with heterogeneous modalities. In *Proc. ICRA*.

[37] Ramanathan, V.; Joulin, A.; Liang, P.; and Fei-Fei, L. 2014. Linking people in videos with their names using coreference resolution. In *Proc. ECCV*, 95–110.

[38] Schmidt, T.; Newcombe, R.; and Fox, D. 2014. DART: Dense articulated real-time tracking. In *Proc. RSS*.

[39] Sener, O.; Zamir, A. R.; Savarese, S.; and Saxena, A. 2015. Unsupervised semantic parsing of video collections. In *Proc. ICCV*, 4480–4488.

[40] Sturm, J.; Stachniss, C.; and Burgard, W. 2011. A probabilistic framework for learning kinematic models of articulated objects. *J. of Artificial Intelligence Research* 41(2):477–526.

[41] Sung, J.; Jin, S. H.; and Saxena, A. 2015. Robobarista: Object part-based transfer of manipulation trajectories from crowd-sourcing in 3D pointclouds. In *Proc. ISRR*.

[42] Tellex, S.; Kollar, T.; Dickerson, S.; Walter, M. R.; Banerjee, A. G.; Teller, S.; and Roy, N. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proc. AAAI*.

[43] Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proc. CVPR*.

[44] Walter, M. R.; Hemachandra, S.; Homberg, B.; Tellex, S.; and Teller, S. 2013. Learning semantic maps from natural language descriptions. In *Proc. RSS*.

[45] Wang, H.; Klaser, A.; Schmid, C.; and Liu, C.-L. 2011. Action recognition by dense trajectories. In *Proc. CVPR*, 3169–3176.

[46] Winograd, T. 1972. Understanding natural language. *Cognitive Psychology* 1:1–191.

[47] Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. ICML*.

[48] Yan, J., and Pollefeys, M. 2006. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In *Proc. ECCV*, 94–106.

[49] Yu, S.-I.; Jiang, L.; and Hauptmann, A. 2014. Instructional videos for unsupervised harvesting and learning of action examples. In *Proc. Int'l Conf. on Multimedia*, 825–828.

[50] Zender, H.; Martínez Mozos, O.; Jensfelt, P.; Kruijff, G.; and Burgard, W. 2008. Conceptual spatial representations for indoor mobile robots. *Robotics and Autonomous Systems* 56(6).