Minimalist Plans for Interpreting Manipulation Actions

Anupam Guha[1], Yezhou Yang[1], Cornelia Fermüller[2], and Yiannis Aloimonos[1]

Abstract— **Humans attribute meaning to actions, and can recognize, imitate, predict, compose from parts, and analyse complex actions performed by other humans. We have built a model of action representation and understanding which takes as input perceptual data of humans performing manipulatory actions and finds a semantic interpretation of it. It achieves this by representing actions as minimal plans based on a few primitives. The motivation for our approach is to have a description, that abstracts away the variations in the way humans perform actions. The model can be used to represent complex activities on the basis of simple actions. The primitives of these minimal plans are embodied in the physicality of the system doing the analysis. The model understands an action under observation by recognising which plan is occurring. Using primitives thus rooted in its own physical structure, the model has a semanticist and causal understanding of what it observes. Using plans, the model considers actions as well as complex activities in terms of causality, compositions, and goal achievement, enabling it to perform complex tasks like prediction of primitives, separation of interleaved actions and filtering of perceptual input. We use our model over an action dataset involving humans using hand tools on objects in a constrained universe to understand an activity it has not seen before in terms of actions whose plans it knows of. The model thus illustrates a novel approach of understanding human actions by a robot.**

I. INTRODUCTION

Humans have an ability to understand meaning in the actions of other humans by observing them. Not only can humans learn complex actions, they can do more than mere imitation of previously observed actions because humans, unlike most AI systems operate in their semantic space. They have the ability to reason and predict future actions of others in real time. To achieve a goal, a human can construct complex plans of actions by having observed only parts of them, having never seen the complete sequence of the right action in the first place. Actions are interesting because infants are sensorimotor agents, and develop abstract concepts later. Certain linguists are of the opinion [1] that even abstract non physical concepts are rooted in physical primitives.

Humans have a robust mechanism to filter and simplify observed behaviour of other humans, as well as a mechanism to compose complex plans out of those observations. We would like to endow robots with a capability of this kind so that they can understand human behaviour and coexist with humans. There are two basic questions one needs to address in order to achieve this goal. **The filtering of sensory data**, is a necessity in understanding the way humans deal with complex actions. The amount of information a human perceives is staggering, and to use it effectively humans can filter out large parts of its perception. How does a human decide what to filter out and what to focus on while observing a complex action in real time, simultaneously understanding it? It has been argued that infants drastically filter out information [2] in order to build schemas or plans with perceptual data. Understanding the **representation of actions** is fundamental to have any kind of mechanism of action understanding. What kind of information is important in Action Representation and how is the information structured? Human activity recognition is an active area of research due to its many potential applications, and currently a lot of different approaches are being followed. Most approaches focus on full body motions and involve observing spatio temporal positions of entities or humans [3], [4], [5] in the universe and learning patterns from them. Some involve trying to model the symbolic representation of sensory data. There have been suggestions [6] that a minimalist generative grammar, similar if not the same as the one which exists for languages, also exists for action understanding, and attempts to understand the grammar of this action-language with the grammar primitives as the trajectories [7] or the objects [8] have been made. However, these grammar models, while addressing how the filtering might be done, do not provide much insight into why the filtering is done in that manner. Also, a Context Free Grammar doesn't provide a lot of explanation as to when an action ends since it does not concern itself much with the causality of the sub-actions.

We suggest a model in which we provide a basis for a semanticist understanding of actions. Firstly, this model attempts to represent complex activities composed of plans of relatively simple action descriptions. Secondly, this model represents these simple actions humans do with their two hands in terms of plans making use of just two primitives, namely 'MOVE' and 'GRASP'. These are what we call "minimal plans". While MOVE and GRASP can be resolved to finer ontologies, this level of abstraction provides a good balance between robustness and adequate information and solves several issues raised above. A plan acts as a description in addition to being a representation. It imparts an understanding of causality. Actions, by their very nature, are goal oriented. There is evidence of action planning in the human brain. Infants have been shown to have the faculty to identify and recreate actions in a hierarchy [9]. We suggest that the grammar of human actions is more than a grammar, and must fundamentally involve planning with a hierarchy of plans arising out of some kind of minimalist plans. Minimal

[1]A. Guha, Y. Yang, and Y. Aloimonos, are from the Department of Computer Science, and [2]C. Fermüller is from UMIACS, University of Maryland, College Park, MD 20742, USA {aguha, yzyang, yiannis} at cs.umd.edu and fer at umiacs.umd.edu

plans also explain 'compositionality', or the ability to understand larger activities based on smaller learned actions, an ability which infants have been observed to have [9]. Most importantly, using minimal plans explains the filtering out of a lot of perceptual data because the goal of a small plan helps the human to focus on particular sub-actions, by observing the changes in the universe due to the visible consequences of earlier actions. This explains the phenomena of 'attention'. This kind of minimal action representation can be used to segment and parse videos robustly due to the very small number of primitives involved. Important also, it gives us a natural way of segmenting the motion sequence in time: a new action starts whenever an object comes in contact with a hand (i.e. the object is grasped)

Our model gives a justification of the two primitives used on a semantic basis which has meaning in terms of the robot. Thus, we attempt to move away from the conventionally functionalist approaches towards semantic mechanisms. A shortcoming most of existing approaches have is that while attempting to understand "meaning" of complex actions, they keep a **functionalist framework** in which the robot doesn't have any understanding of meaning in the symbols it uses. Thus our approach attempts to provide a robot with a very basic form of "intentionality"

II. RELATED WORK

The area of human activity recognition and understanding is one attracting a lot of interest. Generally it can be divided into two types, the visual recognition methods, which comprise of recognition techniques, and the non-visual description methods, which are traditionally functionalist approaches forming representations of actions. A few good surveys of the former can be found in [10], [11], and [12]. There is a lot of motivation for these kind of models as the possibilities of application are immense, especially in areas like HCI, recognition and retrieval, biometrics, in the video domain, and various kinds of image segmentations and classifications. The general method used by most of the visual recognition methods is to learn from a lot of spatio-temporal points what an action looks like and a few works of interest can be found in [3], [4], [5], and [13].Till now most of the focus is in recognizing singular human actions like walking, jumping, running or gait etc. like the work done in [14] and [15]. More complex actions require the usage of HMMs [16] or other parametric approaches for learning the representation as done in works like [17], [18], and [19]. This is the primary thrust of research in computer vision for such recognition techniques.

However, to properly reason about complex actions, semantics are required, and if independent semantics cannot be achieved, at least meaningful syntax is required . Some methods use the concept of regenerative minimalist grammars in analogy with languages. In these methods to relate the objects to actions, HMMs are used in [20] and [21]. Closely related to our work is the work in [22] where an attempt is made to understand complex actions compositionally using a minimalist grammar tree using object detection. This work on action trees based on objects is further developed in [23]. They address the question of actions being compositionally interleaved as well as make an attempt to have action prediction. They use an action grammar structure which is powerful enough to analyse complex action and also resolve interleaved actions. However, the grammar oriented method of most of these works do not address the theme of any independent semantic understanding of actions. That is their systems do not have any actual understanding of the complex actions in terms of what is meaningful to the robot instead of the programmer, or a non-arbitrary theory of where the bootstrapping should be. While our approach to what we think is rudimentary semantic capability may not be a correct analogue of any biological agent, we think that a semanticist justification is necessary for an action understanding model and we attempt to provide one in this paper.

III. DESCRIPTION

In this section we will describe the entire process of understanding actions using minimalist plans by first describing our methodology in collecting data in sub-section III-A, then the concept of minimalist plans and the primitives used in them in sub-section III-B. In subsequent sub-sections III-C, III-D, and III-E we describe how action recognition is done by recognizing which plan is occurring on the basis of primitives and the action alterations. Lastly, in sub-section III-F we talk about automated planning and how it complements this approach.

A. UMD Minimalist action dataset

In order to build our action primitives we require the robot to observe a series of hand-actions performed by a human actor using several different kinds of tools. In addition, we have compiled a new dataset, consisting of 6 actions: A={SLICE, JOIN, MASH, TRANSFER, POUR, STIR} performed by 2 different human actors using 6 common tools: T={knife, ladle, pitcher, ladle, mug, bowl} and 4 other objects[1]: O={tomato, cucumber, bread, cheese}. For the purpose of an accurate tracking of both hands, at the beginning of the clip, a calibration phase is included. In total, there are 10 video clips, which serves as a testbed for the analysis of minimalist action plans.

B. Minimal Plans

To begin with, as mentioned before in the introduction, our model is based on minimal plans. All plans have primitive actions, i.e, actions that cannot be subdivided into anything simpler on some consistent basis, on which the plan gets based. What is the justification for an action to be primitive from the terms of the robot even if those are bootstrapped? We would like that the basis of selecting primitive actions has to be both consistent and non-arbitrary. To address non-arbitrariness first let us ask a question, how does a human get rudimentary knowledge of an action and its consequences on objects? A human does that by imitation [24], that is,

[1]For the rest of the paper, unless specified, the term "object" will be used to indicate both tools and objects. This does gets sorted in the planning stage

it observes as an infant [25] what other humans are doing to same objects under altered conditions and tries to mirror their limb movements to what may be an internal motor [26] . Thus, the most elementary actions seem to be based on the smallest limb movements an actor can produce. We posit that herein lies the basis of rooted meaning. To elaborate we think that ultimately the meaning of an action lies in what it means to the agent's body in terms of the smallest and most fundamental of its physical characteristics. After all the ways in which an agent can modify the universe with its own limbs should be a major part of any comprehension of its universe. Firstly, we only consider actions that can be accomplished by the two hands of a human. As our data shows, we are dealing with a finite domain of actions. Secondly, we abstract all "fine motoric actions" by the fingers due to the limitations of our perceptual apparatus to robustly view those actions. If we allow for these two assumptions, we reach a curious conclusion that to describe all simple actions that can be done by the two hands of a human, we need only two action primitives, namely, that of the action MOVE and the action GRASP where these actions are defined as

- MOVE: Transport the object under observation from one location to the other. The environment or object(s) may or may not alter due to MOVE.
- GRASP: To use the hand, or a suitable robotic analogue, to hold the object under consideration. The opposite of GRASP is UNGRASP. GRASP may or may not alter the environment or objects

Using these two action primitives in our dataset we evaluate six human activities, namely SLICE, JOIN, MASH, TRANSFER, POUR, and STIR. Since these plan scripts have overlapping structures, a condensed unified script is given in Fig. 1. We are considering GRASP and UNGRASP

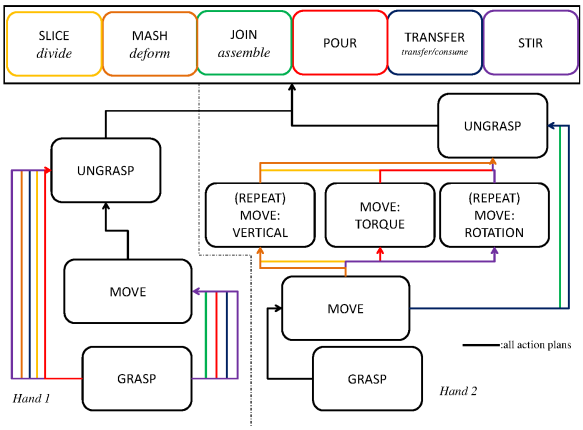

Fig. 1. The six actions of our dataset, SLICE, JOIN, MASH, TRANSFER, POUR, and STIR

as a tuple in the same primitive. Here we clarify a few points. Firstly, we know of course that both MOVE and GRASP have large variety in themselves. [27] considers there are 33 kinds of GRASPs divided hierarchically into 13 groups (and thus corresponding ways to UNGRASP) a human can perform. Similarly, as we will demonstrate later, there are various kinds of MOVE, and currently we are able to detect three kinds of them. Having said that, the basic commonality between all MOVEs and all GRASPs persuades us to use these two as the primitives. Secondly, one may envision other primitives (aside from fine motoric actions which we will not consider), like 'Reach' wherein one reaches towards an object, 'Engage' in which an object is brought to interact with another object etc. and we hold that all these primitives are essentially cases of MOVE. A plan requires a consistent definition of the world, and a finite set of precondition and postconditions in addition to primitive actions. In our minimalist plans, objects are provided with several properties, like "Graspability", "Slicability", "Being a Tool", "Being a Container" etc. Preconditions consist of a combination of existence of objects, their presence in specific locations, and their properties. Postconditions consist of all of these three but can also have what we call "Action Alterations", namely perceivable consequences or changes in the properties of the objects or the environment due to the primitive actions. Each of the six actions mentioned before has an expanded script. In all of these minimal plans there are essentially two scripts occurring simultaneously, one for each of the hands of the human. These are Partial Order Plans in which at times the order in which both the hands do these tasks may be indeterminate. An important point to be stressed is that to infer which plan the robot is observing, it does not need to know all the parameters of the plan to compare it to the scripts it has in its memory. Actually, as we will demonstrate in the next section, if there are a finite number of scripts in consideration, then a few parameters of the plan will be enough to infer which action is being performed. This is a very useful property because one a script of a plan is known, it allows the inference of affordances. Thus, once we know that we are operating in the SLICE plan space, the object the plan is operating upon is a "Slicable" object and the object doing the slicing is a "Slicing Tool". Thus these minimal plans allow an easy way to engage in semantic reasoning. In our simplified action universe we are dealing with only six actions but these actions can be composed into more complex actions. Since these are plans, all our model needs to do is maintain correct causality. Aside from these actions we have worked with, potentially a lot more can be analysed in terms of just our two primitives. After obtaining the scripts of the plans it becomes very easy to determine in an automated manner which script the agent is operating in. However, to learn the new scripts from just observation of human actions with no such bootstrapping in a completely automated manner is beyond the scope of our current work.

C. Plan Detection

As mentioned before, an action is recognized by detecting which plan the agent is currently operating in. For example, the previous Fig. 1 demonstrates that if one hand does only a GRASP followed by an UNGRASP, then it has to be one of the five actions which are not JOIN, as JOIN requires

a movement in both hands. However, if there is a GRASP followed by a MOVE followed by an UNGRASP in both hands, then it cannot be a SLICE or MASH. If one adds in the information of the movements in addition to that of grasping, it is possible to whittle down the possibilities further. Add in the information of locations of objects varying with time and the search space decreases even more. Lastly, if there is still confusion about which action is being done, we have the powerful tool of Action Alteration which determines what consequence the action has done on the environment and/or the object. Knowing all these effectively recognizes the action. As soon as the action is recognised the complete plan is confirmed which in turn allows us to predict affordances, or properties, of the objects involved. To detect the primitive GRASP, we take recourse in the inherent property of moving and grasping, namely that before a GRASP the hand will be in motion while the object will be stationary, and after the UNGRASP it will be the same. However, in the former case the hand will move towards the object whereas in the latter case the hand will move away from the object. Also, a GRASP and the UNGRASP may have none, one, or more MOVEs between them. A property of a hand moving an object is that while in motion both the hand and the object will have relatively similar velocity in magnitude and direction. This will suddenly change as soon as the movement stops and UNGRASP happens. Thus, by making a graph of the relative distances of hands and the objects from each other, and another graph with the relative velocities of the hands and the objects, it is possible to parse out instances of GRASPs, MOVEs, and UNGRASPs by looking for changes in the relative velocities. The method is illustrated in the Fig. 2.

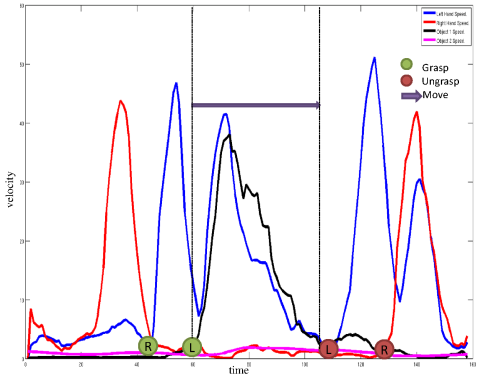

Fig. 2. Parsing out the action primitives using relative velocity

However, a GRASP and an UNGRASP may have more than one sort of MOVE between them. In our minimalist plans we differentiate moves when they undergo a sharp change in property, like a relatively rectilinear motion changing into a periodic up and down motion will count as two MOVEs. We see examples of this happening in four of our actions. To detect these sudden change in movement properties, we plot a highly smoothed graph of the coordinates of the centre of observation of the two objects in the three dimensions which can be parsed to see if the linear motion changes for example to a circular one as shown in Fig. 3 . At this stage in recognition the system has information about the action primitives. But to correctly recognize the action it may need some additional information regarding the consequence of the action. For that there is our procedure of monitoring alterations in objects after actions are performed.

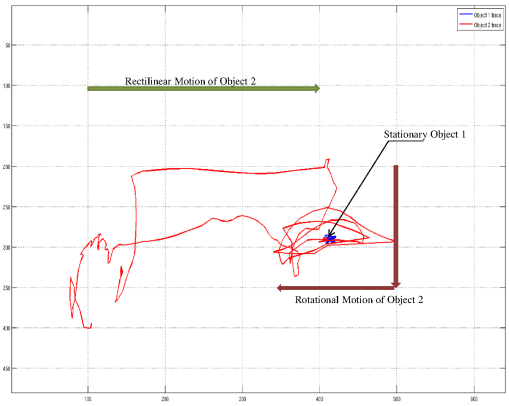

Fig. 3. Using object location to get different MOVEs

D. Hand Tracking and Grasp Type Recognition

We pre-process the dataset using the FORTH hand tracker available[2] [28] which tracks the 3D position, orientation and full articulation of a human hand from markerless visual observations with Kinect input. Currently, we are trying to find if a finer classification of "GRASP" primitive can be obtained using the full articulation and orientation of both hands. While we maintain "GRASP" as semantically atomic, it is useful for future work where we can consider more elaborate means of recognizing which minimal plan we are in. We collect 10 different grasp types following [27] as training data, and extract longitudinal and oblique arches of each finger as features. We further reduce the dimensionality by PCA and then apply k-means clustering to discover the underlying four general types of hand status, 1)REST, 2)FIRM GRASP, 3)DELICATE GRASP (PINCH) and 4)EXTENSION GRASP. (See Fig. 4). To every trail, a naive-bayes classification is used to classify grasp types for each frame.

E. Object Monitoring and Alterations

An object monitoring process is needed to deduce the action primitives. We use the joint segmentation and tracking method presented in [29]. This method combines stochastic tracking [30] with a fixation based active segmentation [31]. The tracking module provides a number of tracked points. The locations of these points are used to define an area of interest and a fixation point for the segmentation, and the color in their immediate surroundings are used in the data term of the segmentation module. The segmentation module segments the object, and based on the segmentation,

[2]http://cvrlcode.ics.forth.gr/handtracking/

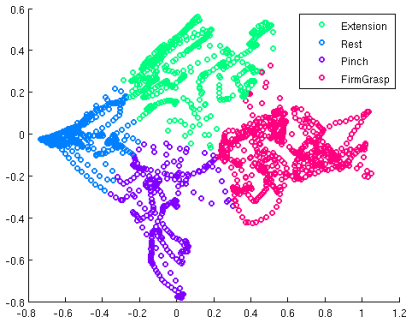

Fig. 4. Four types of grasp and their territories on low dimensional space.

updates the appearance model for the tracker. Fig 5 illustrates the method over time, which is a dynamically closed-loop process. Another crucial pre-condition and post-condition of

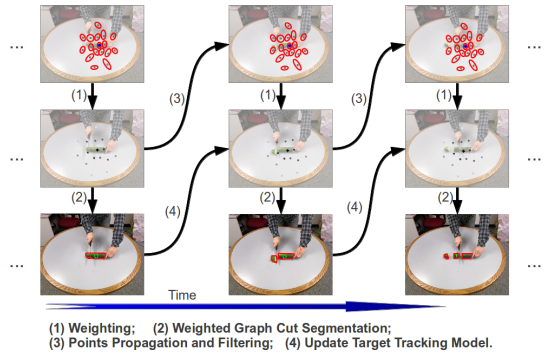

Fig. 5. Flow chart of the proposed active segmentation and tracking method for object monitoring.

human actions is action alterations, aka, the consequence of every action. In the context of our work where we are concerned with how actions change the universe it is necessary to ascertain how they alter the objects they operate on. From their very nature, action alterations can be defined into six primitive categories, 1)DIVIDE, 2)ASSEMBLE, 3)TRANSFER, 4)DEFORM, 5)CREATE and 6)CONSUME. For further details please refer to [29].

Thus knowing the positions of the objects and the hands, parsing the MOVEs and GRASPs from their relative velocities, evaluating what kind of MOVE occurs by the location, and evaluating the action consequence, the complete plan is built and the action recognized.

F. Integration with Automated Planning

Representing actions as minimal plans has another very important advantage, namely that they can be seamlessly integrated with an automated planner. An automated planner can generate a sequence of minimal plan representations for an unknown activity, which can be used to either verify or predict the observed sequence. To a potential robotic platform, using both minimal and complex plans in a hierarchy makes it possible for it to incrementally reason about actions. We converted our primitives in the constrained world to a suitable PDDL [3] representation, and we investigated planning techniques to find a planning algorithm suited to generate sequences of these representations. It should be mentioned here that planning is a PSPACE-complete problem even for simple problems and there is no universal efficient solution for all methods.

The first algorithm which we investigated is FF [32] planning. Its base system architecture uses Enforced Hill Climbing, a forward search engine, [32] as the search algorithm which uses a certain heuristic called relaxed GRAPHPLAN on every state. Enforced Hill Climbing uses a forward searching technique to search through the state space. At every state relaxed GRAPHPLAN estimates the distance from the goal heuristically, and also computes the promising successor states. In addition to this, FF has techniques to avoid wasting time getting to goals needed later which happens where goal orderings are present. This creates problems in using FF in our particular problem. The reason being the action sequences observed had a significant number of primitive actions with indeterminate orders where FF simply fails instead of trying out random combinations of such orders. Thus, while FF is fast where indeterminate orders do not exist it tends to fail gracelessly in the planning problem we are facing. This problem will remain in all state space planners.

The second class of algorithms we investigated were partial order causal link (POCL) planners which search in partial plan space. Partial Order refers to the fact that the plan generated may have primitive actions in indeterminate order. A POCL planner searches for the solution in partial plan space, i.e. it makes a set of partial plans, each with flaws, and then tries to select these partial plans while resolving the flaws. This continues till either a plan without flaws is discovered, in which case it is a success or the partial plan space is exhausted in which case it is a failure. Thus, the problem of ordering or indeterminate primitives never arises. The planner among these best suited to our needs is VHPOP [33] which works where FF fails. VHPOP uses the A^* algorithm to search through plan space and also uses multiple flaw selection strategies concurrently which gives it much more accurate and faster performance than comparative planners. Also, VHPOP has capability for durative planning, something we may incorporate in future work when we incorporate temporal elements into the plans. Tying in our minimalist plans with an automated planner is important for any future application on a robot platform as it gives the robot the option of verification or prediction in addition to directly planning its own actions. The entire semantic framework of the robot operates with knowledge of causality and we approach the ideal of "intentionality".

IV. EVALUATION

As mentioned earlier, our dataset has 10 videos of the six actions which our system is able to recognize. To make certain this recognition is robust, we evaluate it by deliberately inducing artificial noise in the input parses and noting for a particular level of incorrect inputs what will be the accuracy in recognizing the action. By doing this we

[3] http://ipc.informatik.uni-freiburg.de/PddlResources

are able to derive theoretical bounds of accuracy in plan recognition. Also, having built our system which recognises actions by identifying which minimalist plan is occurring we evaluate the robustness of our system by making it observe an activity whose script it does not know of previously, and letting it reason about it in terms of the scripts it knows of.

A. Observing Unknown Activity

Our system observes a complex human activity which is essentially, making a sandwich from a few components like a loaf of bread, a block of cheese and a tomato. This task, aside from being unknown, also has the added complexity of a large amount of objects being simultaneously moved around in the environment making it confusing for the system. The system does not have the sequence of minimal plans representing this entire activity. On observing this activity our system manages to interpret it as a series of interleaved actions of SLICEs and JOINs as it correctly identifies the order of MOVEs and GRASPs and is able to associate them with the consequences for the two actions which are correctly observed. In Fig. 6 the part of the sandwich making process which is one of the JOIN actions is being correctly identified. The two GRASPs and the two MOVEs along with the consequence ASSEMBLE lead to the correct conclusion that this part of the activity is a JOIN. Interestingly, despite the profusion of objects, due to correct tracking of the hand and the relative velocities with objects, it does not make mistakes with deducing the order of activities in making of a sandwich. Also, the entire sequence of actions observed matches nicely with the automated planner's output as expected. The only weak part observed in this experiment was difficulty in tracking the knife in the repeated SLICE actions because of its reflective surface but since we were doing velocity comparisons and since the GRASPs on the knife had been accurately found the knife movements were nevertheless correctly identified.

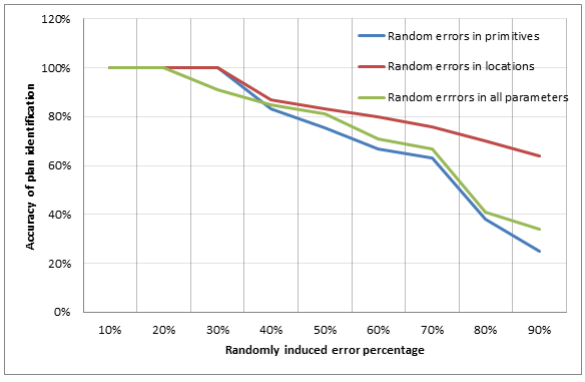

Fig. 7. Accuracy scores for random corruption.

B. Inducing Artificial Noise

In the real world no object detection and segmentation methodology can claim complete accuracy. Occlusion by hand sometimes leads to incorrect segmentation of the objects. Thus our model needs to be robust over imperfect perception. We tested this by first randomly inducing errors in detecting the velocities leading to errors in the detection of the primitives MOVE and GRASP and then we computed the percentages of incorrect plan identification. We did this by finding which plan is the closest to the one being observed. Here, closest is defined as the plan with the least "edit distance" from the six known minimal plans where one edit has the same weight for consequence, primitive and location. We then counted the number of plans correctly identified for each level of error. Then we did the same thing by inducing errors in the location measurement thus making what kind of MOVE we are observing as inaccurate randomly. Thirdly, we induced errors in all of the perceptual inputs simultaneously, the primitives, the locations as well as the consequences, and computed the closest possible plans being identified, and counted how many plans were correctly recognized for what level of induced error. The results of this are in the Fig. 7. Till 40% of all the three categories of errors the accuracy of detecting the correct plan lies in the 80-90% range. The graph shows high resilience even when subjected to $>$60% induced errors in all categories. In the worst case scenarios, inaccuracies in the location doesn't change much of the results because most of the plan is determined by the primitives. Actual performance is much better than these figures because we chose the errors out of uniform error distribution. In reality some aspects of perception, like the consequences and velocity measurement, are generally quite robust so the errors are not evenly distributed. Even with such unfavourable conditions the minimalist planning method exhibits the robustness needed to deal with noise.

V. CONCLUSIONS AND FUTURE WORK

Representing actions by minimalist plans seems to be a useful, versatile, and robust method to do action understanding. Our future work would involve refining the definition of the primitives MOVE and GRASP giving them their own ontologies. This will be done keeping in mind that the robustness of this system depends on the very small amount of primitives. The perceptual apparatus will be made robust to do feature processing in order to better track the objects and be able to observe fine motoric actions by finger mapping. The plans will be used in a feedback loop to develop a prediction and attention mechanism which uses a plan to direct its attention to objects which confirm that the plan is correct thus improving object recognition in the process. Finally, we will work on a method to automatically learn plans in a generative manner by observing human actions on a basis of lesser bootstrapped information. That will allow faithful emulation by a robotic platform of observed human activities.

ACKNOWLEDGMENT

The support of the European Union under the grant Poeticon++ in the Cognitive Systems program, and the support of the National Science foundation under an INSPIRE grant in the Science of Learning Activities program and a grant in the Cyberphysical Systems program are gratefully acknowledged.

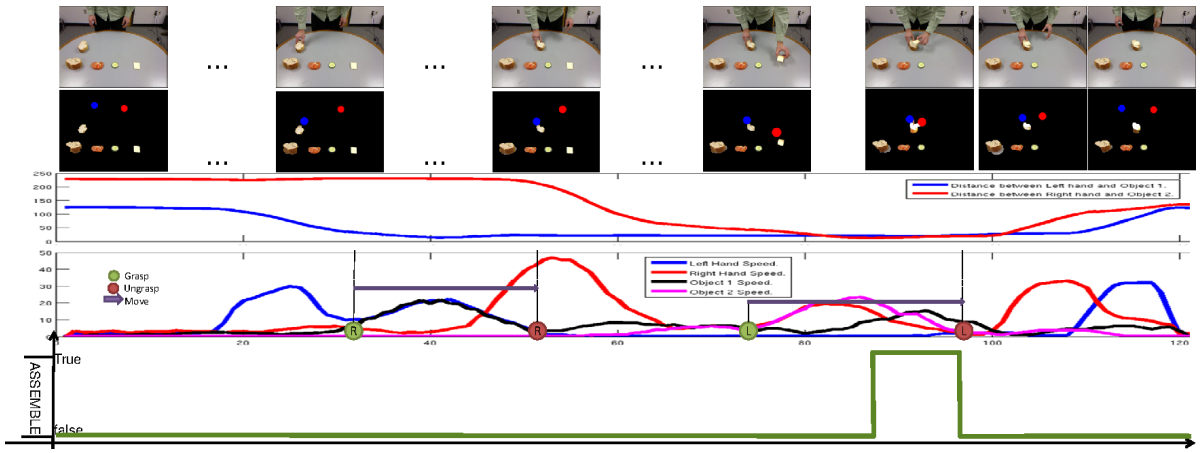

Fig. 6. Illustration of recognizing one of the JOIN actions while making a sandwich. The first row has the results of the hand tracking and object segmentation, the second row has a graph depicting distance of object from hands followed by a row of graphs depicting relative velocities of the objects and hands which detect the GRASPs, UNGRASPs and MOVEs, the last row depicts where the ASSEMBLE alteration is detected.

REFERENCES

[1] G. Lakoff, "Women, fire, and dangerous things," *What categories reveal*, 1987.

[2] J. M. Mandler, "How to build a baby: On the development of an accessible representational system," *Cognitive Development*, vol. 3, no. 2, pp. 113–136, 1988.

[3] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2, pp. 107–123, 2005.

[4] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*. IEEE, 2005, pp. 65–72.

[5] L. Wang and D. Suter, "Learning and matching of dynamic shape manifolds for human action recognition," *Image Processing, IEEE Transactions on*, vol. 16, no. 6, pp. 1646–1661, 2007.

[6] N. Chomsky, *Lectures on government and binding: The Pisa lectures*. Walter de Gruyter, 1993, vol. 9.

[7] Y. A. Ivanov and A. F. Bobick, "Recognition of visual activities and interactions by stochastic parsing," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 8, pp. 852–872, 2000.

[8] D. Moore and I. Essa, "Recognizing multitasked activities from video using stochastic context-free grammar," in *Proceedings of the National Conference on Artificial Intelligence*. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2002, pp. 770–776.

[9] A. Whiten, E. Flynn, K. Brown, and T. Lee, "Imitation of hierarchical action structure by young children," *Developmental science*, vol. 9, no. 6, pp. 574–582, 2006.

[10] T. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer vision and image understanding*, vol. 104, no. 2, pp. 90–126, 2006.

[11] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 11, pp. 1473–1488, 2008.

[12] D. M. Gavrila, "The visual analysis of human movement: A survey," *Computer vision and image understanding*, vol. 73, no. 1, pp. 82–98, 1999.

[13] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," *ECCV*, pp. 650–663, 2008.

[14] J. Ben-Arie, Z. Wang, P. Pandit, and S. Rajaram, "Human activity recognition using multidimensional indexing," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 8, pp. 1091–1104, 2002.

[15] A. Yilmaz and M. Shah, "Actions sketch: A novel action representation," in *CVPR*, vol. 1, 2005, pp. 984–989.

[16] A. Kale, A. Sundaresan, A. Rajagopalan, N. Cuntoor, A. Roy-Chowdhury, V. Kruger, and R. Chellappa, "Identification of humans using gait," *Image Processing, IEEE Transactions on*, vol. 13, no. 9, pp. 1163–1173, 2004.

[17] C. Hu, Q. Yu, Y. Li, and S. Ma, "Extraction of parametric human model for posture recognition using genetic algorithm," in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*. IEEE, 2000, pp. 518–523.

[18] P. Saisan, G. Doretto, Y. Wu, and S. Soatto, "Dynamic texture recognition," in *CVPR*, vol. 2, 2001, pp. II–58.

[19] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *CVPR*, 2009, pp. 1932–1939.

[20] S. Hongeng and R. Nevatia, "Large-scale event detection using semi-hidden markov models," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 1455–1462.

[21] N. Oliver, E. Horvitz, and A. Garg, "Layered representations for human activity recognition," in *Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on*. IEEE, 2002, pp. 3–8.

[22] K. Pastra and Y. Aloimonos, "The minimalist grammar of action," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 367, no. 1585, pp. 103–117, 2012.

[23] D. Summers-Stay, C. Teo, Y. Yang, C. Fermuller, and Y. Aloimonos, "Using a minimal action grammar for activity understanding in the real world," in *Intelligent Robots and Systems, IEEE International Conference on*, 2013.

[24] S. Schaal, "Is imitation learning the route to humanoid robots?" *Trends in cognitive sciences*, vol. 3, no. 6, pp. 233–242, 1999.

[25] M. Bornstein, "A descriptive taxonomy of psychological categories used by infants," *Origins of cognitive skills*, pp. 313–338, 1984.

[26] M. Iacoboni, R. P. Woods, M. Brass, H. Bekkering, J. C. Mazziotta, and G. Rizzolatti, "Cortical mechanisms of human imitation," *Science*, vol. 286, no. 5449, pp. 2526–2528, 1999.

[27] T. Feix, R. Pawlik, H. Schmiedmayer, J. Romero, and D. Kragic, "A comprehensive grasp taxonomy," in *Robotics, Science and Systems Conference: Workshop on Understanding the Human Hand for Advancing Robotic Manipulation*, 2009.

[28] I. Oikonomidis, N. Kyriazis, and A. Argyros, "Efficient model-based 3d tracking of hand articulations using kinect," in *BMVC 2011*, 2011.

[29] Y. Yang, C. Fermuller, and Y. Aloimonos, "Detection of manipulation action consequences (mac)," in *CVPR*, 2013.

[30] B. Han, Y. Zhu, D. Comaniciu, and L. Davis, "Visual tracking by continuous density propagation in sequential bayesian filtering framework," *PAMI, IEEE Transactions on*, vol. 31, no. 5, pp. 919–930, 2009.

[31] A. Mishra, C. Fermuller, and Y. Aloimonos, "Active segmentation with fixation," in *IROS*, 2009.

[32] J. Hoffmann and B. Nebel, "The ff planning system: Fast plan generation through heuristic search," *arXiv preprint arXiv:1106.0675*, 2011.

[33] H. L. S. Younes and R. G. Simmons, "Vhpop: Versatile heuristic partial order planner," *J. Artif. Intell. Res. (JAIR)*, vol. 20, pp. 405–430, 2003.