2011 11th IEEE-RAS International Conference on Humanoid Robots
Bled, Slovenia, October 26-28, 2011

Generalizing Objects by Analyzing Language

Minija Tamosiunaite, Irene Markelic, Tomas Kulvicius and Florentin Wörgötter
Bernstein Center for Computational Neuroscience
University of Göttingen
Friedrich-Hund Platz 1, D-37077
Email: [minija,irene,tomas,worgott]@physik3.gwdg.de

***Abstract*—Generalizing objects in an action-context by a robot, for example addressing the problem: "Which items can be cut with which tools?", is an unresolved and difficult problem. Answering such a question defines a complete action class and robots cannot do this so far. We use a bootstrapping mechanism similar to that known from human language acquisition, and combine language- with image-analysis to create action classes built around the verb (action) in an utterance. A human teaches the robot a certain sentence, for example: "Cut a sausage with a knife", from where on the machine generalizes the arguments (nouns) that the verb takes and searches for possible alternative nouns. Then, by ways of an internet-based image search and a classification algorithm, image classes for the alternative nouns are extracted, by which a large "picture book" of the possible objects involved in an action is created. This concludes the generalization step. Using the same classifier, the machine can now also perform a recognition procedure. Without having seen the objects before, it can analyze a visual scene, discovering, for example, a cucumber and a mandolin, which match to the earlier found nouns allowing it to suggest actions like: "I could cut a cucumber with a mandolin". The algorithm for *g*eneralizing *o*bjects by *a*nalyzing *l*anguage (GOAL) presented here, allows, thus, generalization and recognition of objects in an action-context. It can then be combined with methods for action execution (e.g. action generation-based on human demonstration) to execute so far unknown actions.**

I. Introduction

If you ask your four-year old child: "Cut the sausage with the knife (for eating).", the child will normally have no problems to search for the sausage and the knife on the table and then try to perform the action (hopefully the knife is not too sharp). Recognition of objects in an action context is learned in early childhood. If the child does not like sausages, you may find yourself in the situation where he/she asks: "Can I cut the cucumber instead (because I like it better)?". Here the child has obviously recognized a cucumber as a cuttable object and has generalized the action plan. These – recognition and generalization in an action context – are major cognitive traits of healthy humans already at an early age but no powerful solutions exist so far to transfer these traits to humanoid robots.

In the current study we will present the GOAL[1]-algorithm that links the language- with the image domain for generalizing objects in an action context. Fig. 1 D shows a typical scene with which our robot is confronted. Even if the robot has never seen any of these specific items before, the algorithm allows it to answer the question: "What can be cut with what? (for example: bread with knife) from which an appropriate action could be started. The algorithm is bootstrapped by human language, for example a human uttering a (simple) sentence from where on a generalization process is triggered making use of large text as well as image data-bases for which we use the internet. Similar to some other ontologies (partly used for example in Cyc, see http://www.opencyc.org/), the action (verb!) in the utterance plays here the decisive, class-structuring role around which acting and acted-upon entities are grouped maintaining the semantics of the utterance.

The procedures used in this study for the generalization process are anchored in the theory of human language acquisition. Soon after having learned the first simple phrases, children are able to perform a process called "syntactic bootstrapping" [1]. Essentially this amounts to the inference of the meaning of an unknown word from its grammatical role and the surrounding context of known words. For example: If a child knows the meaning of "fill a cup with water" and then hears the sentence "fill a bilauri[2] with water", it will be able to infer that a bilauri is an object that can be filled (with water). This is a very powerful generalization mechanism that allows young humans to quickly learn the meaning of new words without individual exploration or explicit teaching. Arguably, syntactic bootstrapping leads to the explosion in the language acquisition process that occurs around the age of three in a child [2], [1].

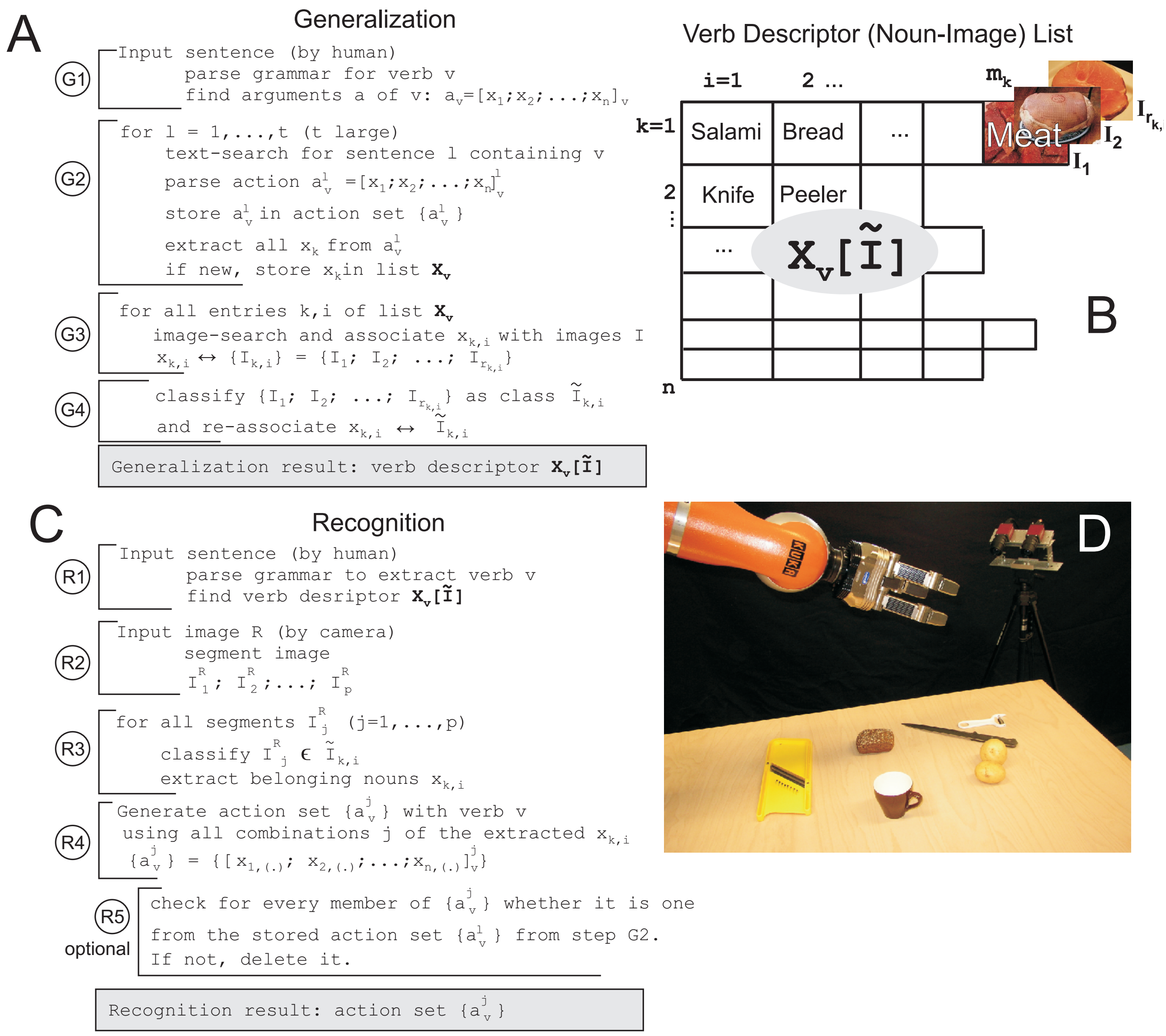
However, a hard problems that remains is how to recognize a potential "bilauri" in the scene. For inexperienced children this amounts to a difficult learning process of structural similarities of the new "fillable object" with other, known ones that some time ago have been filled (or filling has been observed) by the child. Very little is known about how this is achieved. Luckily, in the context of an artificial agent we can circumvent this difficult structural similarity problem and rely on a memory-based image classification process again using internet data as "memory". Different from a young child, a robot can directly access massive amounts of memory using internet data, or data from other large data-bases, and this way it can recognize different objects as shown below.

The goal of the current paper is, thus, to describe the novel GOAL-algorithm that allows the learning of large action classes. In some earlier studies we have shown how an action

[1] Generalizing objects by analyzing language

[2] This means "glass" in Swahili, assuming that most readers have not heard this word before.

978-1-61284-868-6/11/$26.00 ©2011 IEEE



```
(G1) Input sentence (by human)
         parse grammar for verb v
         find arguments a of v: $a_v=[x_1;x_2;...;x_n]_v$

(G2) for l = 1,...,t (t large)
         text-search for sentence l containing v
         parse action $a_v^l = [x_1;x_2;...;x_n]_v^l$
         store $a_v^l$ in action set $\{a_v^l\}$
         extract all $x_k$ from $a_v^l$
         if new, store $x_k$in list $\mathbf{X}_v$

(G3) for all entries k,i of list $\mathbf{X}_v$
         image-search and associate $x_{k,i}$ with images I
         $x_{k,i} \leftrightarrow \{I_{k,i}\} = \{I_1; I_2; ...; I_{r_{k,i}}\}$

(G4)     classify $\{I_1; I_2; ...; I_{r_{k,i}}\}$ as class $\tilde{I}_{k,i}$
         and re-associate $x_{k,i} \leftrightarrow \tilde{I}_{k,i}$

Generalization result: verb descriptor $\mathbf{X}_v[\tilde{\mathbf{I}}]$
```

```
(R1) Input sentence (by human)
         parse grammar to extract verb v
         find verb desriptor $\mathbf{X}_v[\tilde{\mathbf{I}}]$

(R2) Input image R (by camera)
         segment image
         $I_1^R; I_2^R; ...; I_p^R$

(R3) for all segments $I_j^R$ (j=1,...,p)
         classify $I_j^R \in \tilde{I}_{k,i}$
         extract belonging nouns $x_{k,i}$

(R4) Generate action set $\{a_v^j\}$ with verb v
         using all combinations j of the extracted $x_{k,i}$
         $\{a_v^j\} = \{[x_{1,(.)}; x_{2,(.)}; ...; x_{n,(.)}]_v^j\}$

(R5) optional
         check for every member of $\{a_v^j\}$ whether it is one
         from the stored action set $\{a_v^l\}$ from step G2.
         If not, delete it.

Recognition result: action set $\{a_v^j\}$
```

Fig. 1. A, C) Pseudo-code overview of the GOAL-algorithm for generalization and recognition. For description of the variables see text; B) Example of list $\mathbf{X}$ for the verb "slice" and visualization how the noun entries are supplemented by corresponding images. This 3D-list is called verb descriptor $\mathbf{X_v}[\tilde{\mathbf{I}}]$. D) Example of the robot setup for the recognition phase.

data base for a robot can be learned from observation [3], [4]. Similar results from human demonstration exist from several groups [5], [6], [7]. The current paper allows linking such action data bases to self-generated object classes from which an execution process can be triggered.

We will first describe the algorithm and its results and in the discussion section embed this into the current state of the art.

II. Overview of the Algorithm

The GOAL algorithm consists of a *generalization* and a *recognition* process. The key issue is how to generalize from one suggested action into a complete action class. Fig. 1 A,B summarizes the main components of the generalization process (steps G1, G2) and makes the link to the recognition process (steps G3, G4). The bottom part of the figure (panel C) addresses the recognition process. The algorithm is based on the English language but similar principles can be used with other languages, too. The process is described for one example (one verb).

A. Generalization

In the first step (G1, parsing) a human enters a simple sentence similar to the examples above. Such a sentence should consist of a single phrase with one verb v. Verbs have a certain valency, hence they are preceded or followed by

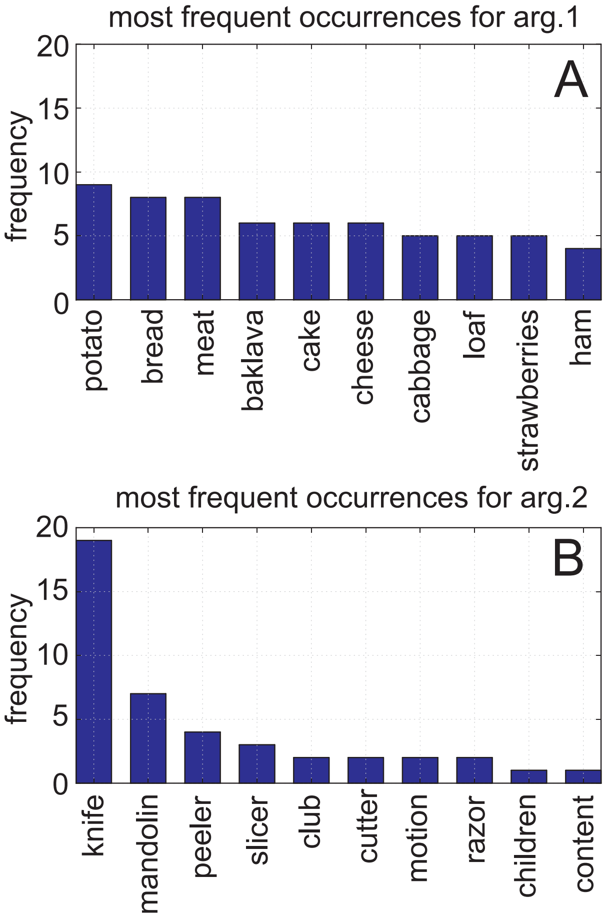


Fig. 2. Verb argument generalization results obtained through text-mining.

words, called arguments, to which the verb v refers[3]. This list of arguments is first extracted by a grammar parser rendering a list $[x_1; ...; x_n]_v$, where n is the valency index (usually $n \leq 3$). For our "cutting" example from the introduction we find $x_1 = sausage$ and $x_2 = knife$. Essentially $[x_1; ...; x_n]_v$ captures an action, thus we call it *action descriptor* a_v.

The next step (G2, ontology kernel) performs text-based generalization. To this end any large text data base can be employed. The data base is searched for sentences l in which the verb v occurs and we store for this specific verb each action descriptor instantiation $a_v^l = [x_1; ...; x_n]_v^l$. Let l be the 5th found sentence, then this descriptor could, for example, read: $a_{cut}^5 = [salami, knife]$. All a_v^l are then again stored in a set $\{a_v^l\}$ for later use. Action sets, such as $\{a_v^l\}$, are the formal way we use to describe the above mentioned action classes, which the robot is supposed to generate. Furthermore we extract all argument instantiations x_k^l. We do this for many sentences l and combine all found argument instantiations into a non-recurring set writing them into list $\mathbf{X_v} = [x_{k,i}]$ (see Fig. 1 B). For example things to cut are: $x_{1,(.)} = \{$salami, cucumber, carrot, bread, wood,...$\}$. Things with which you can cut are: $x_{2,(.)} = \{$knife, cleaver, saw,...$\}$. Individual rows of this list are normally of different length, hence $m_i \neq m_j \; for \; i \neq j$. If sorted by occurrence frequency one finds that mostly less than 10 arguments are common. Thus, this represents the kernel of a verb-based ontology, where verbs are encoded as classes (or concepts or categories, which are often used synonymously) in addition to nouns.

Step three (G3, picture book) links generalization to recognition. For this we perform for every instantiation $x_{k,i}$ an image search using any large image data-base. This way we attach to every noun instantiation of each argument a list of images $I_{k,i} = \{I_1; ...; I_{r_{k,i}}\}$. This way we have effectively created a large picture book about (for example) "What can be cut by what?". If correctly built then all possible combinations of images from x_1 with those from x_2 represent a potentially valid cutting action. As shown below, this picture book allows an agent to actually recognize in the visual scenery potential objects to be used for a given action. Fig. 1 B provides a summary of the data structure that has been generated by steps G1-G3.

In step four (G4, classification) an image classification procedure is used to extract and represent the complete image class by its class label $\tilde{I}_{k,i}$. This is done for every instantiation $x_{k,i}$. A quick estimate[4] shows that picture book storage does not produce any capacity problem for more than 1000 verbs hence going far beyond the requirements for generalization of (for example) the corpus of "Simple English" [8], [9].

This ends the generalization process. This way we have finally received a verb descriptor $\mathbf{X_v}[\tilde{\mathbf{I}}]$ by which potential actions are encoded.

B. Recognition

The recognition part is depicted in Fig. 1 C. Step R1 of this process is simple. Either one gives the robot a command, for example: "Please, cut something!", or the robot has by some other means an existing plan for a certain action. At this stage the system needs to extract the verb v and the corresponding verb descriptor $\mathbf{X_v}[\tilde{\mathbf{I}}]$.

[3] Avalent: It rains. Monovalent, $n = 1$: Paul sleeps., Divalent, $n = 2$: Betty kicks the ball, Trivalent, $n = 3$: The robot cuts the salami with a knife, etc. Valency can change for the same verb. For example, "cutting" can be tri- or divalent (see Introduction), but this does not affect the algorithm.

[4] A rough storage estimate can be obtained as follows: Let us consider 1000 divalent verbs. Reasonable generalization for every day use should be obtained with about 20 instantiations per valency. Good image representation - according to our experience – needs 200 pictures of $0.5MB$ per instantiation each. Thus, we get $1000 \times 2 \times 20 \times 200 \times 0.5MB = 4.0TB$. Given that many instantiations frequently re-occur, because the same noun can be used in many different sentences, it is fair to divide this number at least by 5, getting less than $800GB$ of required storage for about 1000 verbs without any compression. When using feature-based storage (e.g. with SIFT features) this is reduced by a factor larger than 10, leading to just $80GB$ of required storage. Simple English contains about $2000 - 3000$ words [8], [9] hence far less than 1000 verbs. Thus, generalization should be possible far beyond Simple English use.

Step R2 takes a camera image R and segments it into those regions that represent single objects. Any advanced computer vision image segmentation procedure can be used here [10], [11], [12]. This way, we receive subparts $I_j^R, j = 1, \cdots, p$ of the original image.

In step R3 we test every image subpart I_j^R whether it belongs to any image class $\tilde{I}_{k,i}$ from step G4. Once classification is successful we use the mutual correspondence $\tilde{I}_{k,i} \leftrightarrow x_{k,i}$ (step G4, above) and we extract from the list $\mathbf{X_v}$ all corresponding noun instantiations $x_{k,i}$, which belong to the image subparts. For example, let I_j^R be an image of a salami. Then it will be classified as $I_j^R \in \{\tilde{I}_{1,1} \leftrightarrow x_{1,1}\}$, where $x_{1,1}$ represents the noun "salami" (see Fig. 1 B), which is one possible first argument ($k = 1$) in the sentence: "Cut the $x_{1,1} = salami$ with a $x_{2,i} = unspecified$.", where we still have to search for a possible second argument.

Step R4 takes all recognized instantiations $x_{k,i}$ and creates action descriptors a_v^j of potentially executable actions where j is one possible combination of arguments: $a_v^j = [x_{1,(.)}; x_{2,(.)}; ...; x_{n,(.)}]_v^j$. We store all a_v^j in the action set $\{a_v^j\}$. All combinations of instantiations are allowed. This way sentences like: "Cut the salami with a knife.", "Cut the wood with a saw." but also "Cut the wood with a knife." are obtained.

Step R5 is optional. We can rank the occurrences of actions suggested by the recognition process $\{a_v^j\}$ and also check which of them are represented in the earlier stored action set from the generalization process $\{a_v^l\}$. Likely, sentences like "Cut the wood with a knife" have never been found in the data-base and can this way be ruled out. It may, however, also be of interest to allow for the complete action set without this deletion step, as without deletions the set $\{a_v^j\}$ might contain novel valid actions not observed in the text search.

III. Processing an Example

The goal is to allow the robot to arrive at useful action plans when looking at the scene in Figure 1 D.

We use an example similar to above: "Slice the sausage with a knife". We use "slice" and not "cut" as it is more specific making the text search easier. For steps G1 and G2 we take the Natural Language Processing Toolkit from Python. The sentence has the following grammatical structure, expressed by regular expressions:

$$\{< VB.* >< DT >? < JJ > * < NN.* >< IN > \\ < DT >? < JJ > * < NN.* >\},$$

where $< VB.* >$ is a verb of any tense and voice, $< DT >?$ a 0 or 1 determiner ("the", "a"). $< JJ > *$ is any number of adjectives, $< NN.* >$ represents any declined noun and $< IN >$ exactly one preposition. For this sentence $< JJ > *$ remains empty and we get two instances of $< NN.* >$, which represent $x_1 = sausage$ and $x_2 = knife$.

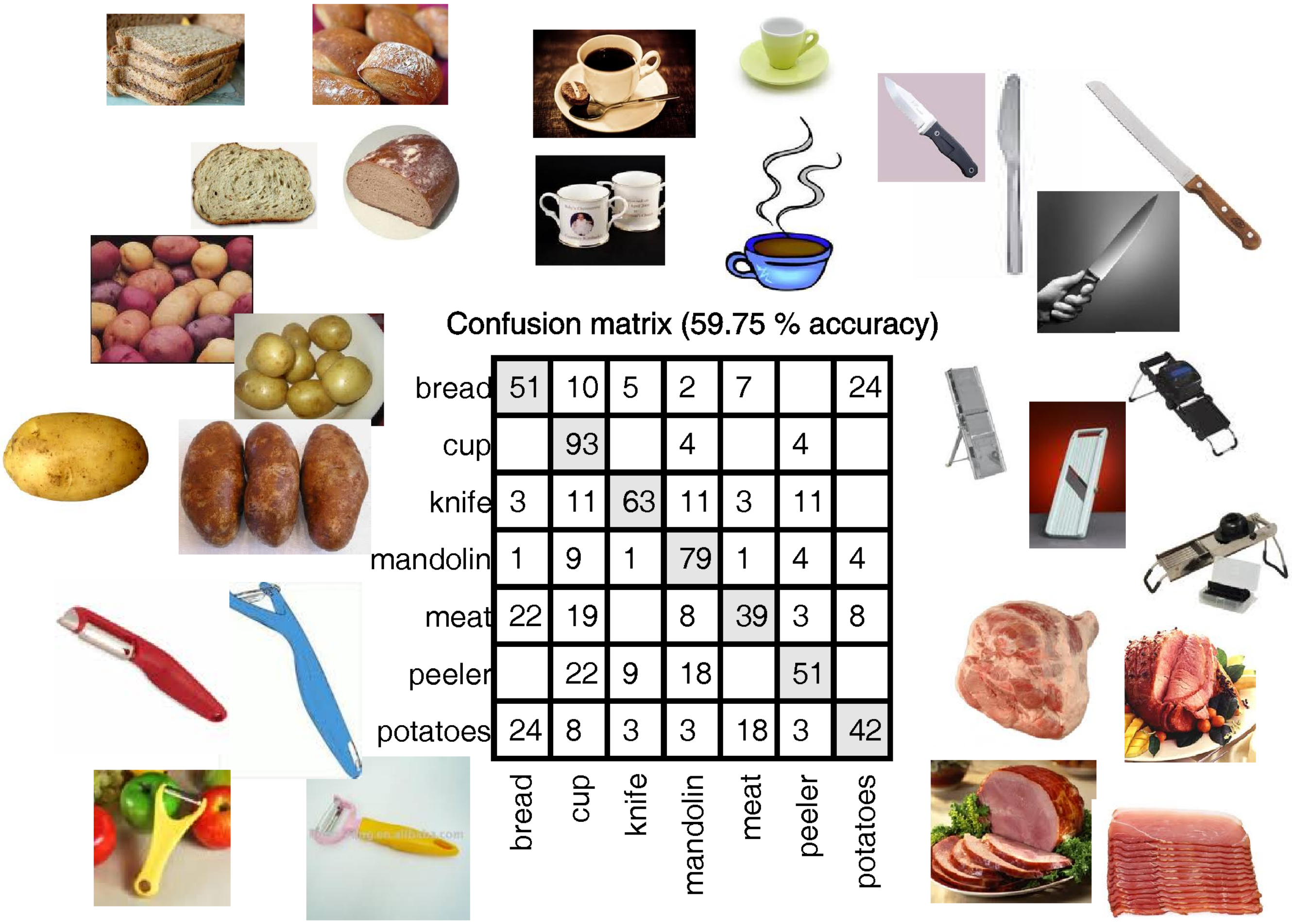
Step G2 performs internet-based text search using the following strings: "Slice the $< JJ > * < NN.* >$ with a knife" and "Slice the sausage with a $< JJ > * < NN.* >$". We use the first 200 web-pages found for each of the two strings and extract occurrence frequency histograms for the nouns (Fig. 2). The first 10 entries for each argument $x_{(1,2)}$ are used as instantiations $x_{(1,2),i}$. One can see that the most frequent ones are indeed appropriate. Errors occur for example with noun modifiers ("children" really meant "children knife" in this entry). To solve these problems better parsers (e.g. predicate-argument parsing [13]), could be used.

Steps G3 (picture book) and G4 (classification) requires image search as well as image data-base cleaning. For this we use the VLFeat Package [14] and perform feature-based classification by the use of the Support Vector Machine available in VLFeat. Visual features are pyramid histogram of edge orientations gradients (PHOG) and pyramid histogram of visual words (PHOW), which are based on the well known SIFT (scale invariant feature transform) framework [15], [16]. In Fig. 3 we show the confusion matrix for 7 examples from a total of 250 pictures per class corresponding to words from parts of the set belonging to argument 1 (sausage&similars) and argument 2 (knife&similars). In addition we have included a set of images of cups, which is not an argument of the verb "slice". The results demonstrate that reasonably well distinguishable image classes $\tilde{I}$ can be obtained and that other objects (cups) can also be discriminated. This concludes the generalization part of the algorithm.

For testing we have recorded 9 uncluttered images from different views similar to the setup in Fig. 1 D with as-yet unknown objects that potentially belong to the two arguments $x_{(1,2),i}$ of the verb "slice". Fig. 4 shows one example from those 9 images used for analysis. For step R2, we used a variant of our standard image segmentation algorithm [12] including a foreground-background segmentation stage. As uncluttered scenes were used this renders 8 separate image sub-parts $I_{(1,2,...,8)}^R$. Bounding boxes are drawn in the Figure only for graphical reasons. For step R3, we use the classifier trained in step G4 and classify subparts I_1^R as "potatoes", I_2^R as "cup", I_3^R as "mandolin", I_4^R and I_5^R as "unknown", I_6^R as "bread", I_7^R as "knife" and I_8^R as "unknown". This assessment is based on a voting scheme using all 9 different camera views. Voting is required, because state of the art image classification do on average not reach more than 60% recognition rates (like our method). The peeler, we used, seems to deviate too much from the examples with which the classifier was trained and could not be recognized. Such a behavior is sometimes to be expected given the current state of the art in image classification.

Step R4 is straightforward. It performs the re-combination of the different recognized arguments into valid sentences, which are: slice a (1) bread with a knife, (2) bread with mandolin, (3) potato with knife, (4) potato with mandolin. These are potentially meaningful actions, but - as discussed above - some are not really useful (bread with mandolin).

In Step R5 we compare the suggested actions with the ones



Confusion matrix (59.75 % accuracy)

	bread	cup	knife	mandolin	meat	peeler	potatoes
bread	51	10	5	2	7		24
cup		93		4		4	
knife	3	11	63	11	3	11	
mandolin	1	9	1	79	1	4	4
meat	22	19		8	39	3	8
peeler		22	9	18		51	
potatoes	24	8	3	3	18	3	42

Fig. 3. Image classification results and examples from picture book.

found in G2 on the internet and we find that, indeed, the sentence: "Slice (a/the) bread with (a/the) mandolin." has not been observed. This allows ruling this action out.

Alternatively the machine would have to try the action out, or it would – like a child – have to ask: "Can I slice a bread with a mandolin?". It is important to realize that the GOAL-algorithm does indeed provide us with largely reasonable early guesses about valid actions, thereby being able to guide an agent through the vast richness of perceived objects. But final, conclusive evidence about the validity of an action cannot be expected.

IV. Discussion

In the current study we have suggested a bootstrapping algorithm which leverages from the structure of language combining it with image classification for the generalization and recognition of large action classes. It is based on minimal supervision and can, thus, be included in dialogue-based robot interaction. Extension of the robot's data base for action classes is induced by human speech. The potential for dialogue is discussed below (Error Handling and Extensions) as a means to efficiently extend the capabilities of the algorithm.

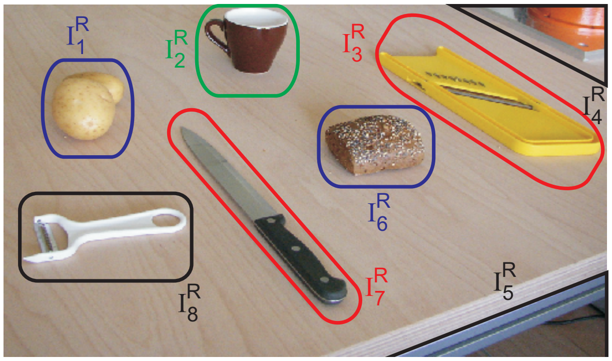


Fig. 4. Recognition results for a scene observed by a robot.

The strength of the GOAL-framework lies, thus, in this interactive aspect but especially in its conceptional simplicity: the sequence of the individual bootstrapping steps (G1-G4; R1-R4) is straightforward. These two aspects form the basis of the algorithm, where it is important to note that the individual components (language and computer vision analysis) can be continuously improved making the algorithm open-ended. As soon as available, more complex computer linguistic methods can be used to parse more complex sentences, or more complex scenes can be analyzed with improved computer vision. In addition, the GOAL-framework can be linked to robot execution (discussed below).

A potential weakness is that many individual steps have to be performed which can produce inappropriate results and currently the framework has been tested only with a small set of examples. In general, as with all cognitive algorithms, exhaustive testing can only ”come through use”. There are no benchmarks existing and correctness can only be assessed by the robot’s human counterpart. This will take several years.

At the moment we can assess potential errors of the different components from a small set of experiments and find that steps G1, G2 and R1 are performed with few errors only. These arise for example from a wrong treatment of noun-modifiers (”children’s knife”) or other similar effects. Many such errors can be removed by using a more advanced parser. In general we will in the future continue to use only simple sentences and advanced computer linguistics can deal with this level of complexity essentially in an error-free way. It is important to note that the complexity of the analyzed sentences does not necessarily improve the bootstrapping process as such.

Computer vision analysis is potentially more problematic. For steps G3, G4, and R3, we obtain currently a classification accuracy of about 60%. The problem can be mitigated by introducing thresholds and enabling classifier to produce ”I don’t know” output. In a robot application false positives are not acceptable but ”dunno” is usually ok because the robot can continue to explore the scene (active vision!), getting better views of the unrecognized object thereby being able to finally classify it.

A. State of the Art

Massive efforts exist since a few years to make the internet available to robots and to adapt it for their specific needs (see http://www.roboearth.org/). The underlying approaches are quite different from what we are trying to achieve here, because GOAL combines image classification with text-based generalization to enrich a robot’s action space by forming large categories of potentially executable combinations of objects with actions. Image classification and text mining are heavily researched. Common-sense knowledge from targeted databases has already been applied in robotics [17], [18]. But as far as we see the literature background, we come to the conclusion that the specific combination of text and image base mining we are proposing has been very little investigated in the context of robotics [19].

Many of the individual components that we have used for this algorithm are now common standard (like simple parsers for text mining, SIFT-related features for image processing, SVMs for classification, etc.) and require no further discussion. At the moment we have not yet paid too much attention to any better and more optimal choice of such components. This notwithstanding, this paper shows that even with simple ones already promising results are obtained. Clearly the underlying problems of general text mining or general image classification remain exceedingly difficult, but we think that the core of our algorithm will only benefit from future improvements.

The main scientific link, which has spawned this idea is the above discussed syntactic bootstrapping [1] in language acquisition (see Introduction). Another link exists to content dependent image classification and retrieval [20], [21]. Though general content-based image retrieval is still largely unresolved, in limited domains like our manipulation scenario it is possible to obtain good enough results for developing experimental applications.

Essentially, our approach can be best understood as partly supervised generation of an action-based ontology (specifically, verb-based ontology) using automatic and conjoint text-image mining. Much work has been done concerning ontologies, where most of them are noun-based [22], [23] and often rely on expert knowledge [24]. We rely on recent trends to extract knowledge from texts automatically [25]. In our system ontology kernels arise automatically linking a verb to its possible arguments, where the relevant set of verbs is provided through human supervision. Obviously statements like ”Cut the wood with a knife.” and other even more nonsensical ones are possible, too. But the strength of the algorithm is that such sentences could either be ruled out by step R5 (deletion) or, due to the link to robotics, such statements could be tried out (by trying this out with the machine) or dialogue could be engaged asking whether or not an action is feasible, or as a forth possibility, more advanced reasoning could be employed after enriching the ontological kernels with additional information (”wood is hard”). We see a major strength of the algorithm that it is open to such extensions. Furthermore, using a verb-based structure makes combined text-image mining feasible as verb-arguments (usually nouns!) can be image-searched for while searching for verbs is less ”pictorial” and will in general not work.

B. Algorithmic Extensions: Error Handling and Extensions by Dialogue

When generalizing the sentence: ”Fill the cup with water”, the system is troubled by the fact that pictures of water and its similars classify in a way which is useless for a filling action. This is due to the fact the pictures of water usually show rivers, lakes, or splashing water whereas for a filling action we would have to search for water in a container. The algorithm will, however, produce a class from the first, useless set of examples. Thus, when asked to perform a ”filling action”, water&similars will not be found and the agent should respond with: ”I do not recognize water or similar items?” This problem could be solved by combined search, for example using ”water fill” as string, or - more interestingly - by again using GOAL: To the robot’s question the supervisor should answer by the sentence: ”Water is contained in buckets.” The GOAL-algorithm can generalize this sentence in the same way, where after step G2 an ontological link emerges. We receive: ”Fill the cup&similars with water&similars” and ”water&similars is contained in buckets&similars”. Thus, with any reasonably good logic reasoner we can infer: ”Fill the cup&similars with water&similars, which is contained in buckets&similars”. The ontology kernels, which are created by GOAL (G2) are verb-centered, and many verbs can take the same arguments. Thus, action chains naturally emerge. This opens the framework up for extending the kernels into a true ontology on which reasoning mechanisms can operate.

Other problems exist, for example syntactic ambiguities (”Cut the meat with the onions and the carrots.”), which need to be resolved by better semantic understanding for which tools like WordNet [26] might help but especially this could be achieved by human-robot dialogue, which may in addition be used to resolve other language- as well as image recognition problems. The aspect that already the generation of the ontology is guided by a human makes such an extension into more elaborate human-robot dialogue quite appealing.

C. The Link to Execution

Much work has been done in the last years to arrive at robot-compatible descriptions of human actions. Mostly this has been achieved by learning from demonstration [27], [28] and several frameworks exist which allow encoding trajectories and required poses to perform a certain action [29], [30]. One of the main remaining difficulties is to associate appropriate objects to an action. The algorithm presented in this study tries to address this problem as it allows finding the arguments of a verb and the associated objects in a visual scene by which it can be linked to execution and first results for simple actions exist [31].

Acknowledgment

The research leading to these results has received funding from the European Communitys Seventh Framework Programme FP7/2007-2013 (Specific Programme Cooperation, Theme 3, Information and Communication Technologies) under grant agreement no. 270273, Xperience.

References

[1] J. Trueswell and L. Gleitman, “Learning to parse and its implications for language acquisition,” in *Oxford Handbook of Psycholinguistics*, G. Gaskell, Ed. Oxford: Oxford University Press, 2007, pp. 635–656.

[2] F. Tracy, “The language of childhood,” *Am. J. Psychol.*, vol. 6, no. 1, pp. l07–138, 1893.

[3] E. E. Aksoy, A. Abramov, F. Wörgötter, and B. Dellen, “Categorizing object-actions relations from semantic scene graphs,” in *Proc. 2010 IEEE Int. Conf. Robotics and Automation*, 2010, pp. 398–405.

[4] E. E. Aksoy, A. Abramov, J. Dörr, N. KeJun, B. Dellen, and F. Wörgötter, “Learning the semantics of object-action relations by observation,” *Int. J. Robot. Res.*, vol. 30, pp. 1229–1249, 2011.

[5] H. Kjellström, J. Romero, and D. Kragić, “Visual object-action recognition: Inferring object affordances from human demonstration,” *Comput. Vis. Image Und.*, vol. 115, no. 1, pp. 81–90, 2011.

[6] M. Sridhar, G. A. Cohn, and D. Hogg, “Learning functional object-categories from a relational spatio-temporal representation,” in *Proc. 18th Europ. Conf. Artificial Intelligence*, 2008, pp. 606–610.

[7] I. Laptev and P. Prez, “Retrieving actions in movies,” in *Proc. IEEE Int. Conf. Computer Vision*, 2007.

[8] H. Kucera and W. N. Francis, *Computational analysis of present-day American English*. Providence, RI: Brown University Press, 1967.

[9] W. N. Francis and H. Kucera, *Frequency analysis of English usage*. Houghton Mifflin, Boston, 1982.

[10] P. F. Felzenszwalb and D. P. Huttenlocher, “Efficient graph-based image segmentation,” *Int. J. Comput. Vision*, vol. 59, pp. 167–181, Sep. 2004.

[11] D. Comaniciu and P. Meer, “Mean shift: A robust approach toward feature space analysis,” *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 24, pp. 603–619, 2002.

[12] A. Abramov, T. Kulvicius, F. Wörgötter, and B. Dellen, “Real-time image segmentation on a gpu,” in *Facing the multicore-challenge*, R. Keller, D. Kramer, and J.-P. Weiss, Eds. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 131–142.

[13] M. Surdeanu, S. Harabagiu, J. Williams, and P. Aarseth, “Using predicate-argument structures for information extraction,” in *Proc. 41st Ann. Meet. Association for Computational Linguistics*, 2003, pp. 8–15.

[14] A. Vedaldi and B. Fulkerson, “VLFeat: An open and portable library of computer vision algorithms,” 2008. [Online]. Available: http://www.vlfeat.org/

[15] A. Bosch, A. Zisserman, and X. Munoz, “Image classification using random forests and ferns,” in *Proc. 11th Int. Conf. Computer Vision*, 2007.

[16] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vision*, vol. 60, pp. 91–110, 2004.

[17] R. Gupta and M. J. Kochenderfer, “Common sense data acquisition for indoor mobile robots,” in *Nat. Conf. Artificial Intelligence*, 2004, pp. 605–610.

[18] L. Kunze, M. Tenorth, and M. Beetz, “Putting people’s common sense into knowledge bases of household robots,” in *Proc. 33rd Ann. German Conf. Advances in Artificial Intelligence*. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 151–159.

[19] M. Tenorth, U. Klank, D. Pangercic, and M. Beetz, “Web-enabbled robots,” *IEEE Robotics & Automation Magazine*, vol. 18(2), pp. 58–68, 2011.

[20] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, “Learning object categories from google’s image search,” in *Proc. 10th Int. Conf. Computer Vision*, vol. 2, 2005, pp. 1816–1823.

[21] R. Datta, D. Joshi, J. Li, and J. Z. Wang, “Image retrieval: Ideas, influences, and trends of the new age,” *ACM Comput. Surv.*, vol. 40, pp. 1–60, 2008.

[22] A. Gangemi, R. Navigli, and P. Velardi, “The ontowordnet project: Extension and axiomatization of conceptual relations in wordnet,” in *Proc. CoopIS/DOA/ODBASE’03*, 2003, pp. 820–838.

[23] P. Buitelaar and P. Cimiano, Eds., *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*. IOS Press, 2008, vol. 167.

[24] H. Liu and P. Singh, “Conceptnet: a practical commonsense reasoning toolkit,” *BT Technol. J.*, vol. 22, pp. 221–226, 2004.

[25] J. Gordon, B. Van-Durme, and L. Schubert, “Learning from the web: Extracting general world knowledge from noisy text,” in *Proc. WikiAI 2010*, 2010.

[26] G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross, and K. Miller, “WordNet: An online lexical database.” *Int. J. Lexicograph.*, vol. 3(4), pp. 235–244, 1990.

[27] A. Billard, S. Calinon, and F. Guenter, “Discriminative and adaptive imitation in uni-manual and bi-manual tasks,” *Robot. Auton. Syst.*, vol. 54, pp. 370–384, 2006.

[28] T. Asfour, P. Azad, F. Gyarfas, and R. Dillmann, “Imitation learning of dual-arm manipulation tasks in humanoid robots,” *Int. J. Hum. Robot.*, vol. 5, pp. 183–202, 2008.

[29] S. Schaal, P. Mohajerian, and A. Ijspeert, “Dynamics systems vs. optimal control–a unifying view,” *Prog. Brain Res.*, vol. 165, pp. 425–445, 2007.

[30] S. M. Khansari-Zadeh and A. Billard, “BM: An iterative method to learn stable non-linear dynamical systems with gaussian mixture models,” in *Proc. 2010 IEEE Int. Conf. Robotics and Automation*, 2010, pp. 2381–2388.

[31] E. E. Aksoy, B. Dellen, and F. Wörgötter, “Execution of pushing actions with semantic event chains.” in *Humanoids 2011*, 2011, in press.