Visual Commonsense for Scene Understanding Using Perception, Semantic Parsing and Reasoning

S. Aditya[1] Y. Yang[2] C. Baral[1] C. Fermuller[2] Y. Aloimonos[2]

[1]School of Computing, Informatics and Decision Systems Engineering
Arizona State University

[2]Department of Computer Science
University of Maryland, College Park

AAAI Spring Symposium Series, 2015

Outline

Outline

Ultimate Challenges in Artificial Intelligence

- Image Understanding.
  - Automatic image annotation (latest using RNN+CNN)
- Agents playing games against humans.
  - DeepMinds deep reinforcement learning playing Atari games.
  - And of course, IBM Watson playing Jeopardy.

Ultimate Challenges in Artificial Intelligence

- Image Understanding.
  - Automatic image annotation (latest using RNN+CNN)
- Agents playing games against humans.
  - DeepMinds deep reinforcement learning playing Atari games.
  - And of course, IBM Watson playing Jeopardy.

Our goal: Image Understanding.

Ultimate Challenges in Artificial Intelligence

- Image Understanding.
  - Automatic image annotation (latest using RNN+CNN)
- Agents playing games against humans.
  - DeepMinds deep reinforcement learning playing Atari games.
  - And of course, IBM Watson playing Jeopardy.

Our goal: Image Understanding.

Erm. Ok! But what is new?

Motivation

"Well, humans are able to deal with cluttered scenes. They are able to deal with huge numbers of categories. They can deal with inferences about the scene: **What if I sit down on that? What if I put something on top of something?** These are far beyond the capability of todays machines."

— Michael Jordan to Spectrum Magazine, Oct 2014

Deep learning is good at certain kinds of image classification. "What object is in this scene?"

Some Negative Examples(RNN+CNN-Karpathy et. al. 2014)

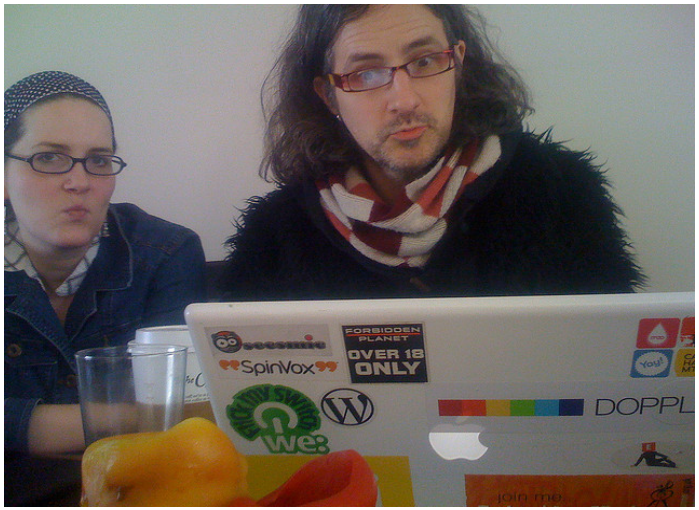


Figure : (a) A man is sitting on a bench with a dog, (b) A woman holding a teddy bear in front of a mirror.

Figure : A bunch of bananas are hanging from a ceiling

In search of Intelligence

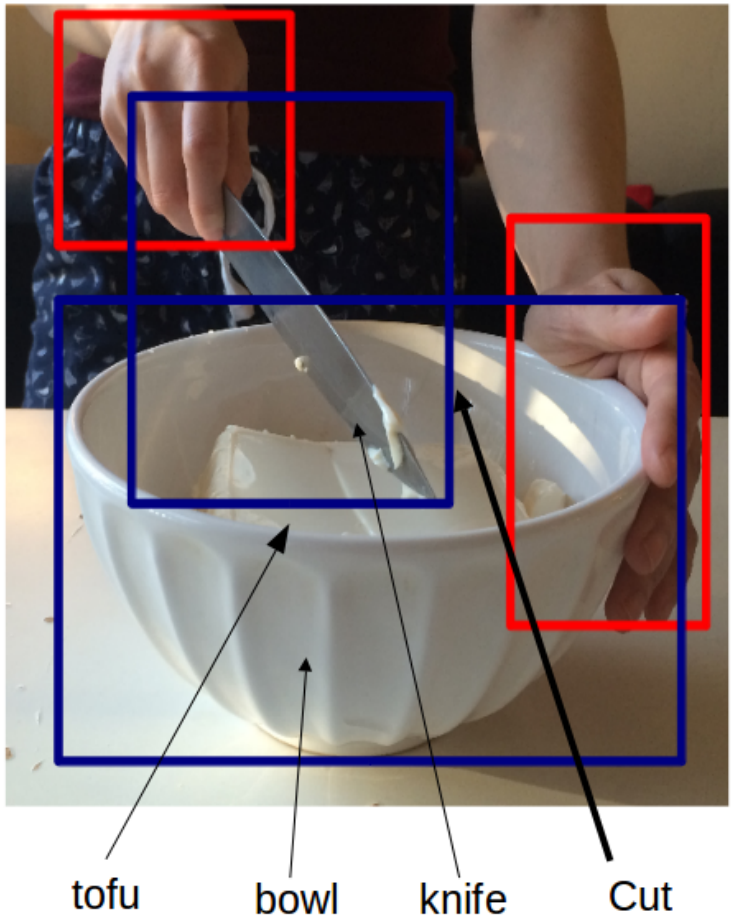
- RNN+CNN State-of-the-art+great accuracy. But why "such" meaningless results?
- How to you replicate humanoid intelligence then?
- This was our motivation.
- This paper provides a humble set of ideas in this direction.

What can we achieve with Common-sense?

The Tofu Example

- Take the figure for example.
- We detect knife, cut, bowl. Say, we do not detect tofu.
- In general, we detect object,action,actedUpon.
- Not possible.
- Use commonsense.
- Knife cutting something inside bowl i.e. tofu.

Cutting Tofu



Use of Common-sense

- Correctly recognize a single action.
- *Is there any other use? There is...*

Use of Common-sense

- Correctly recognize a single action.
- *Is there any other use? There is...*
- Make conclusions regarding finer aspects of an activity.
- Inference of higher-level activities given action constituents.
- Learn the decomposition of activities to individual actions from visual examples. And many more...

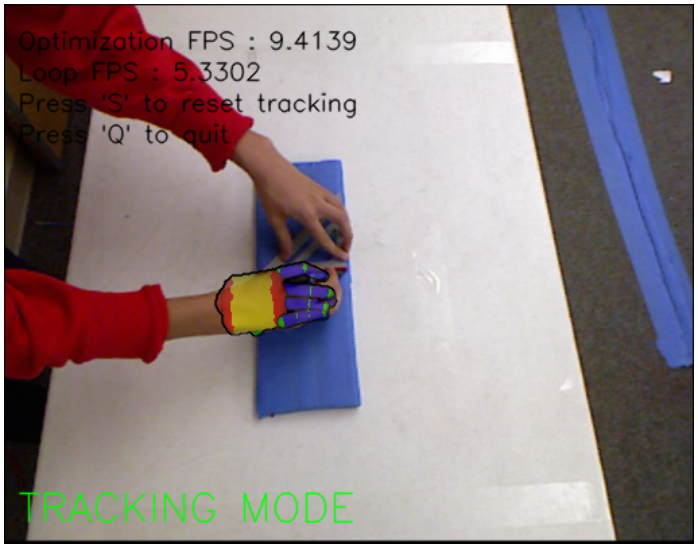
Outline

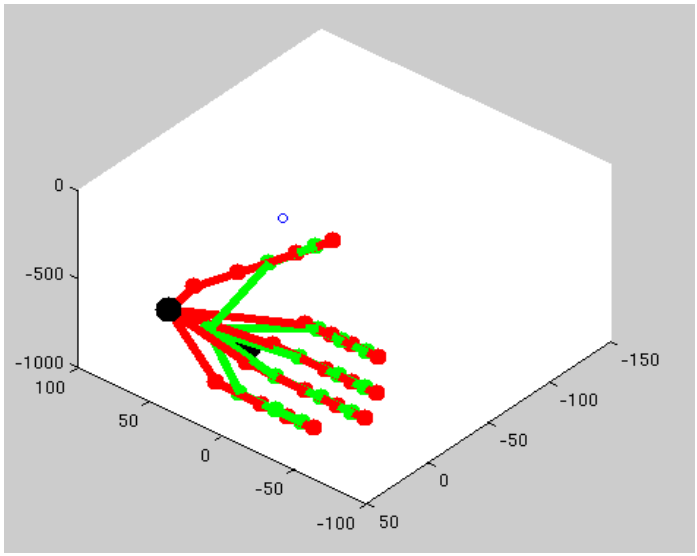
Related Works

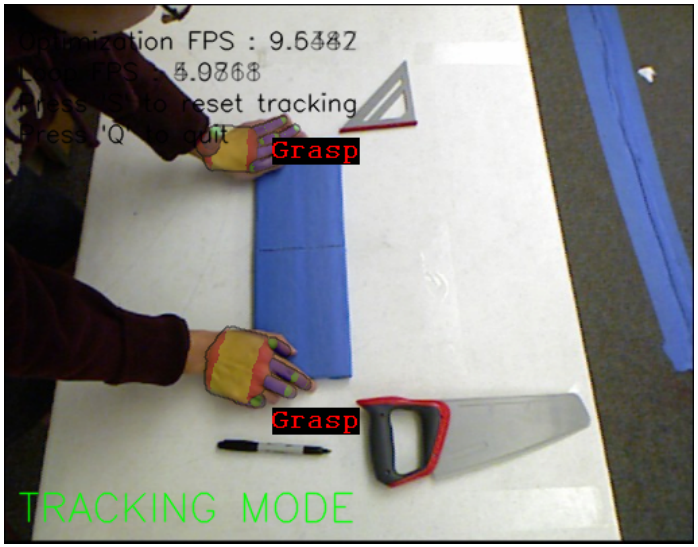
- Action understanding:
  - feature based approaches: [5] Dollr, P. Rabaud, V. Cottrell,G. and Belongie, S. 2005. *Behavior recognition via sparse spatio-temporal features*; [8] Laptev, I. 2005. *On space-time interest points*.
  - More complex actions: [3] Chaudhry, R. Ravichandran, A.Hager, G.and Vidal, R. 2009. *Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions.* [2] Aksoy, E.E. and Abramov, A. and Dörr, J. and Ning, K. and Dellen, B. and Wörgötter, F., *Learning the semantics of object-action relations by observation.*
- Reasoning beyond vision:
  - [12] Dan Xie, Sinisa Todorovic, and Song-Chun Zhu. 2013. *Inferring dark matter and dark energy from videos.* [7] Joo, Weixin Li, Francis F Steen, and Song Chun Zhu. 2014. *Visual persuasion: Inferring communicative intents of images.* [10] Pirsiavash, Carl Vondrick, and Antonio Torralba. 2014. *Inferring the why in images.*

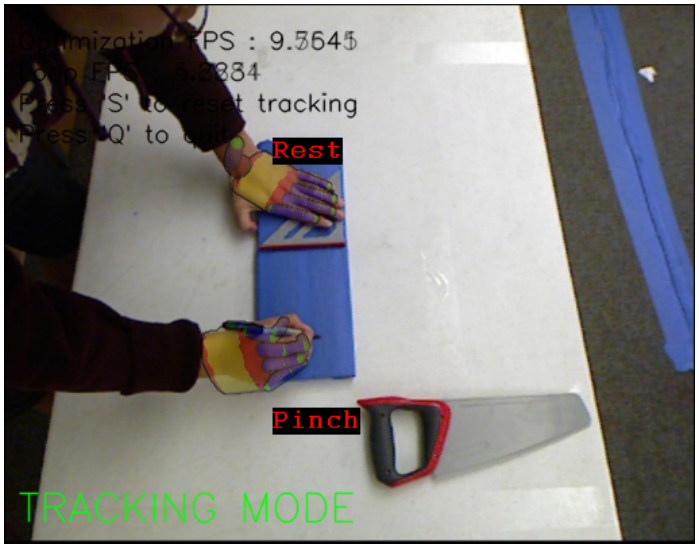
Outline

Hand Tracking and Grasp Type Recognition





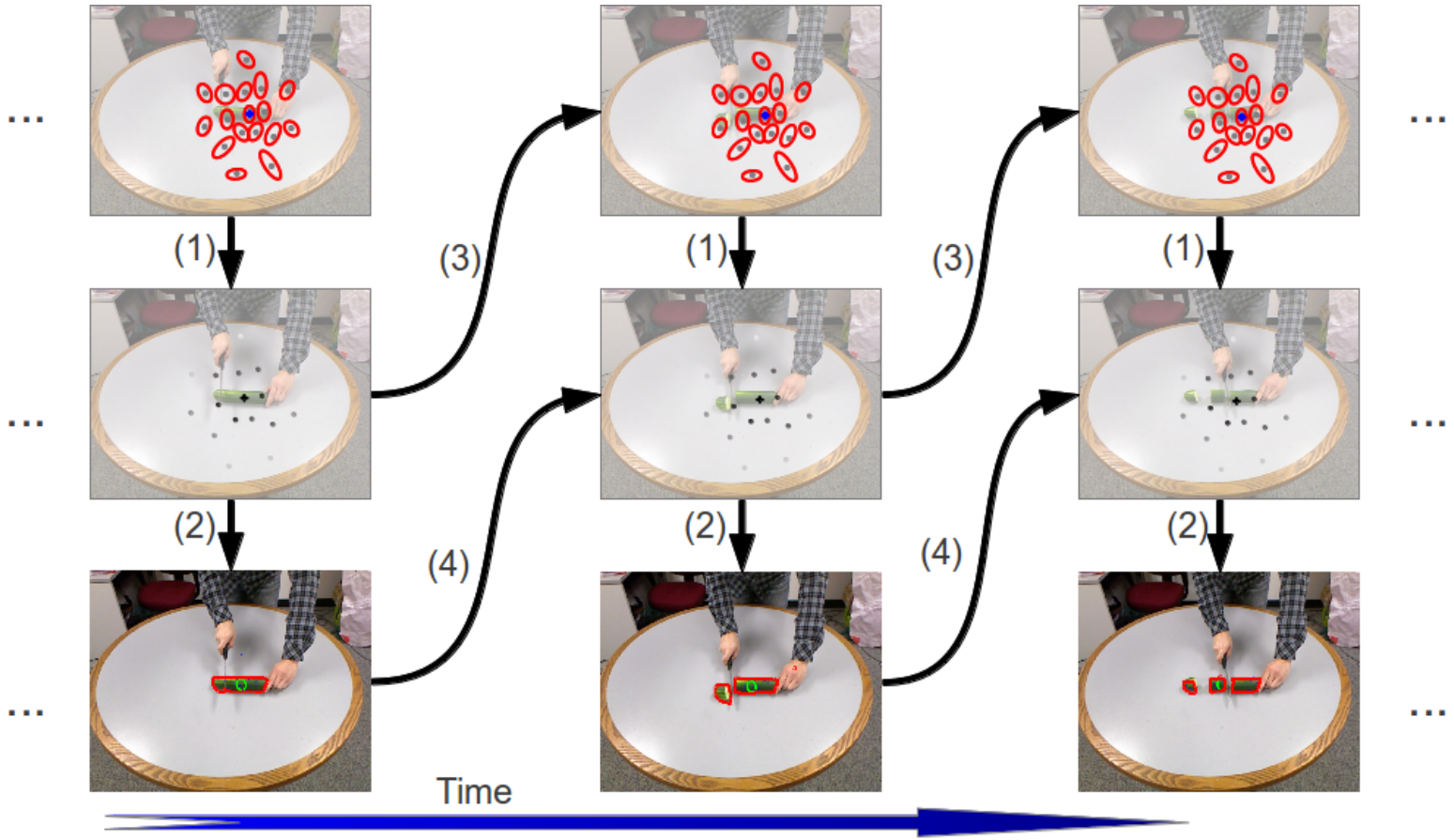




Articulated hand-tracking using Kinect: Oikonomidis, I. 2011[9]

Outline

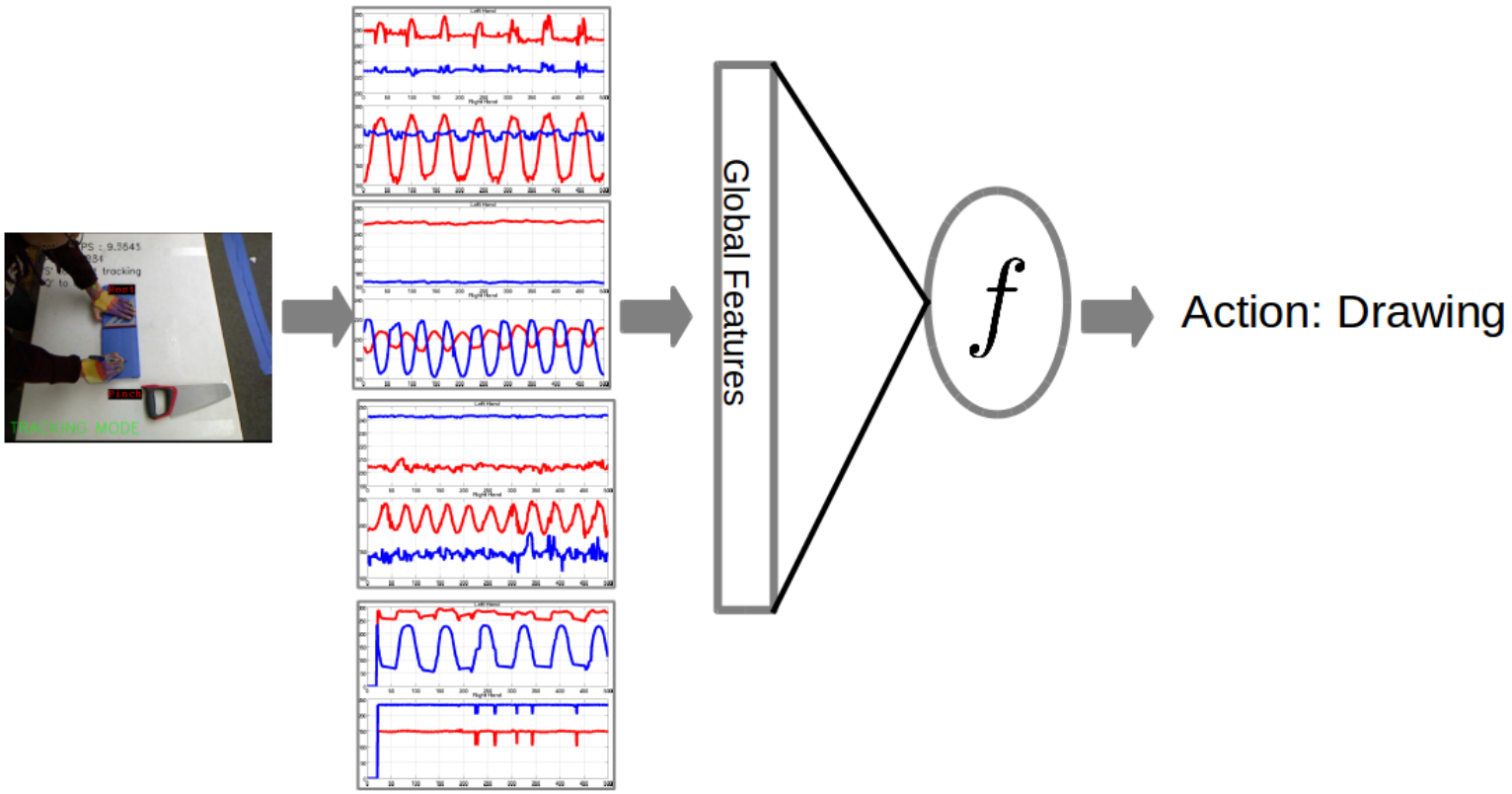
Object Monitoring and Recognition



Detection of manipulation action consequences (MAC), Yang, Y. 2013 [13]

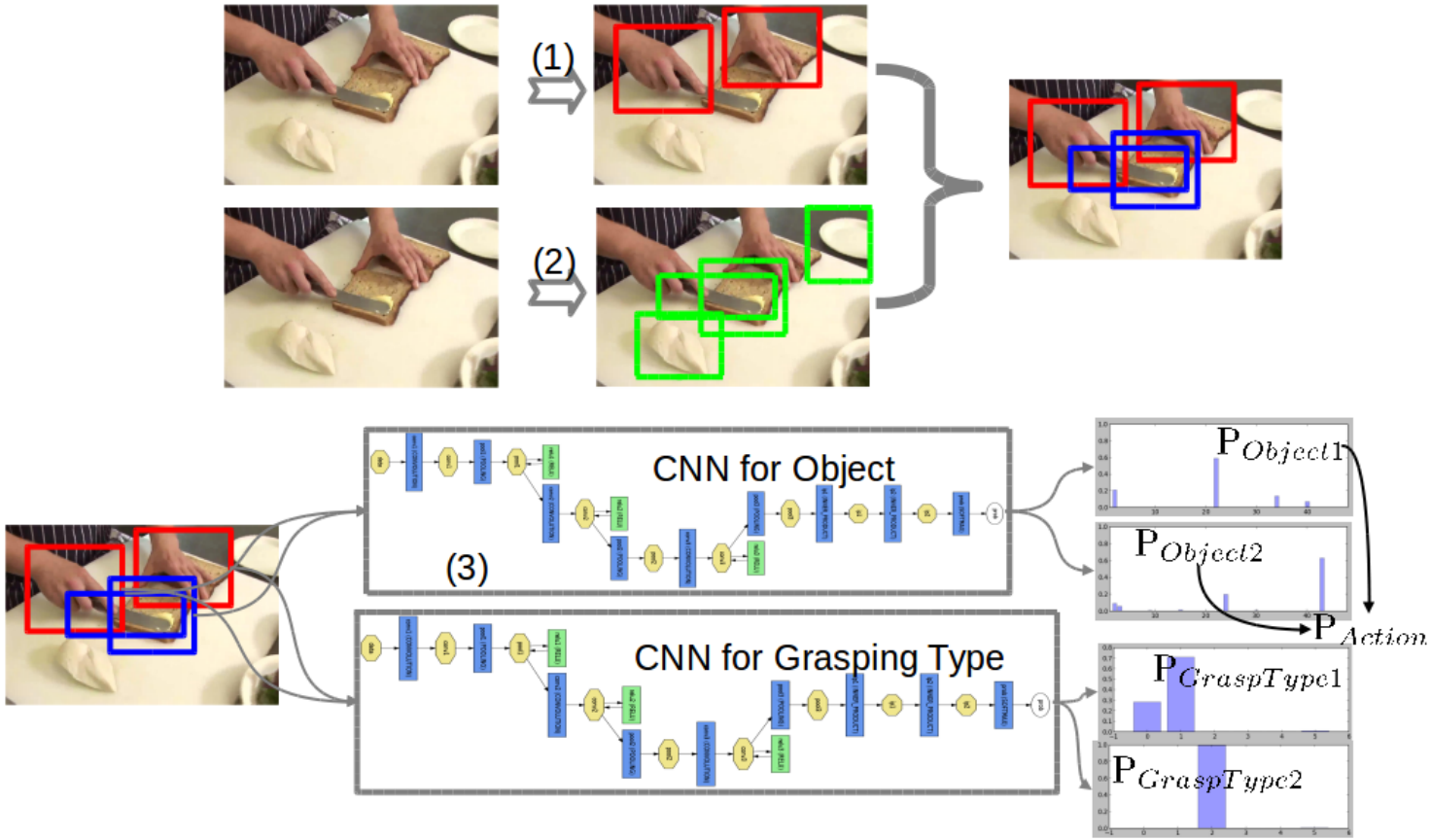
Outline

Trajectory Based Action Recognition



Towards a Watson that sees: Language-guided action recognition for robots. Teo, C. 2012 [11]

Towards unconstrained visual inputs



Mittal, A. 2011 [1]; Cheng, M. 2014[4]; Jia, Y. 2013[6]

Outline

Output from Vision

Output predicates

$appears(put, tofu, bowl, 1).$
$appears(cut, knife, bowl, 2).$
$appears(< action >, object1, object2, timestamp)$
or
$occurs(< action >, object1, object2, < start >, < end >).$

Improving Visual Perception

Reason

- Take the knife,cut,bowl example.
- *How do we automate it?*
- Theory: Knife is an artifact. Bowl is an artifact. Artifact cutting artifact is abnormal.
- *Voila!! Are we done?*

ASP Code

```
appears(put,tofu,bowl,1).
appears(cut,knife,bowl,2).
artifact(knife).
artifact(bowl).
holds(in(X,Y),T+1):- occurs(put,
   X,Y,T).
holds(F,T+1):- holds(F,T), not
   nholds(F,T+1).
nholds(F,T+1):- holds(F,T), not
   holds(F,T+1).
occurs(A,S,O,T):- appears(A,S,O,T),
    not ab(A,S,O,T).
ab(cut,S,O,T):- artifact(S),
    artifact(O).
occurs(cut,S,O,T):- appears(cut,
    S,O,T), ab(cut,S,O,T),
    holds(in(O,O),T)).
```

Whats more...

- Not yet.
- We still need to know that knife is an artifact.

Whats more...

- Not yet.
- We still need to know that knife is an artifact.
- Developed own semantic parser.
- Consults WordNet superclasses.
- Determines artifact or not.
- Available at www.kparser.org

Outline

Finer Aspects of Activities

- Input: a set of facts about a video of marking a line.

```
occurs(grasp1,lefthand,plank,50,85).
occurs(grasp2,lefthand,ruler,95,280).
occurs(align,ruler,plank,100,168).
occurs(grasp3,righthand,pen,130,260).
```

- Questions we can ask:
  1. Which hand is being used in aligning the ruler?
  2. Which hand is used in drawing?
  3. Is the ruler aligned when the pen is drawing on the plank?

Answering Questions...

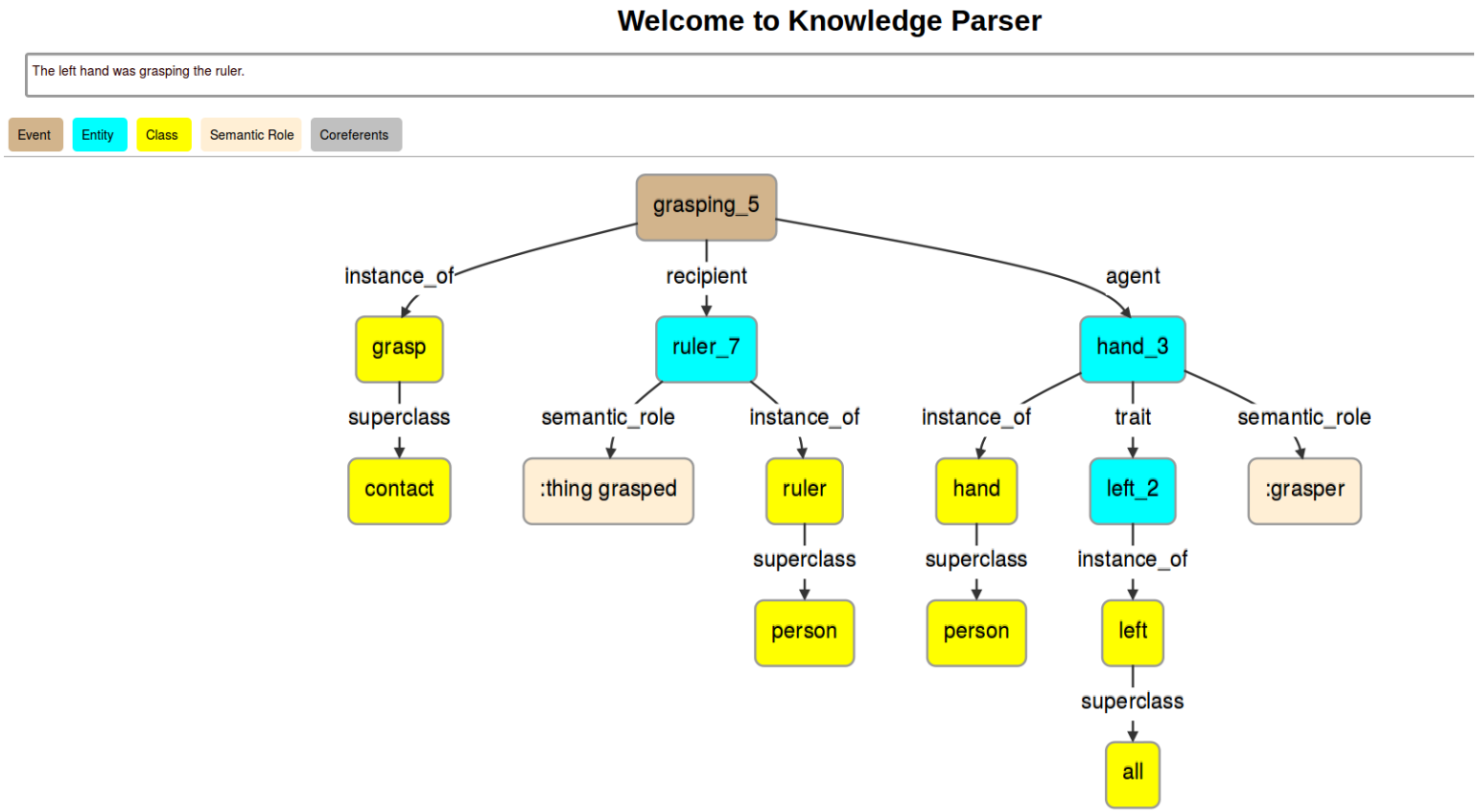
- For Question 1 and 2:
  - If occurs(A1,X,Y,T1,T2) , THEN X and Y was used in A1.
  - If X was used in action A1, H was used in action A2(grasp) and H is a grasper , then hand H was used in grasping X.
- For question 3:
  - We need rules for effect of actions, inertia axioms and the question that is asked.

Answering Questions...

- For Question 1 and 2:
  - If occurs(A1,X,Y,T1 Y was used in A1.
  - If X was used in action A1, H was used in action A2(grasp) and H is a grasper , then hand H was used in grasping X.
- For question 3:
  - We need rules for effect of actions, inertia axioms and the question that is asked.

An information we get from K-parser

Knowledge from K-parser



ASP Program Example

```
start(grasping,T1) :-occurs(grasp1,X,Y,T1,T2).
end(grasping,T2) :-occurs(grasp1,X,Y,T1,T2).
start(grasping,T1) :-occurs(grasp2,X,Y,T1,T2).
end(grasping,T2) :-occurs(grasp2,X,Y,T1,T2).
start(grasping,T1) :-occurs(grasp3,X,Y,T1,T2).
end(grasping,T2) :-occurs(grasp3,X,Y,T1,T2).
start(aligning,T1) :-occurs(align,X,Y,T1,T2).
end(aligning,T2) :-occurs(align,X,Y,T1,T2).
start(drawing,T1) :-occurs(draw,X,Y,T1,T2).
end(drawing,T2) :-occurs(draw,X,Y,T1,T2).
```

```
holds(aligned,T2+1) :-occurs(align,X,Y,T1,T2).
holds(drawn,T2+1) :-occurs(draw,X,Y,T1,T2).
holds(F,T+1) :- holds(F,T), not nholds(F,T+1).
nholds(F,T+1) :-nholds(F,T), not holds(F,T+1).
no :- start(drawing,T1), end(drawing,T2),
     T1 < T, T < T2, not holds(aligned,T).
yes :- not no.
```

Outline

Reasoning in Activity Recognition

Two main aspects:

- Common-sense knowledge about general structure of activities and use the knowledge to recognize a new activity.
- Common-sense in learning the general structures of activities from very few examples.

A Humble Formulation

- Activity: hierarchy of small actions.
  - That is can be broken down to smaller activities, can be further broken down to individual actions.
- Temporal Notion.
- We use following predicates to completely define activity:
  - $component(SA, X, Y, A)$ - short action SA with parameters X and Y is component of A
  - $startsb4ov(SA1, SA2)$ - start time of SA1 is before the start time of SA2, they have some overlap, but SA2 is not a subinterval of SA1.
  - $subinterval(SA1, SA2)$ SA2 is subinterval of SA1.
  - $before(SA1, SA2)$ SA1 ends before SA2.

Our input becomes...

```
component(g1,grasper1,plank,mark).
component(g2,grasper1,ruler,mark).
component(g3,grasper2,ruler,mark).
component(align,ruler,plank,mark).
component(draw,pen,plank,mark).
before(g1,g2).
subinterval(align,g2).
subinterval(g3,g2). subinterval(draw,g3).
startsb4ov(align,g3).
neq(grasper1,grasper2).
grasper1 in {righthand, lefthand}
grasper2 in {righthand, lefthand}
```

Activity Recognition using ASP

- Given a new example, all you need to check now is that all the components satisfy and the order is maintained.
- Given a set of *occurs* predicates $\Rightarrow$ get *component/before/startsb4ov/subinterval* predicates $\Rightarrow$ if all satisfy, THEN its marking a line.
- Hard constraints are assumed here.
- Next, we give one notion to generalize.

Learning Activity Structures

- Hard constraints are not good.
- There could be many ways of doing the same thing.
- Probabilistic formulation might be needed.
- A whole lot of complexity.

What is the simplest thing you can do?

Generalize based on Timeline

- Idea comes from following notion:
  - Its hard to get too many examples of a same video.
  - No Big Data here.
  - Quite hard to analyze patterns.
- Our problem becomes
  1. Just one video i.e. one sample and learn the activity.
  2. Do not forget to generalize.

Simple Solution

- Generalize the components without prior knowledge
- Optimize the time-constraints.
- We introduce possible versions of the predicates given an example:
  - ⋆ Pcomponent, Pstartsb4ov,poverlap, psubinterval, pbefore.

Possible versions

```
pcomponent(g1,lefthand,plank,mark).
pcomponent(g2,lefthand,ruler,mark).
pcomponent(g3,righthand,ruler,mark).
pcomponent(align,ruler,plank,mark).
pcomponent(draw,pen,plank,mark).
pbefore(g1,g2).
pbefore(g1,align).
pbefore(g1,g3).
pbefore(g1,draw).
pbefore(align,draw). psubinterval(align,g2).
psubinterval(g3,g2). psubinterval(draw,g3).
pstartsb4ov(align,g3). psubinterval(draw,g2).
```

Again, a Simplisitic Solution

If we assume the the knowledge that:

- lefthand and righthand are graspers,
- Ruler is an aligning artifact, and
- Pen is a writing instrument.

We get the resulting predicates:

```
component(g1,grasper1,plank,mark).
component(g2,grasper1,aligner1,mark).
component(g3,grasper2,aligner1,mark).
component(align,aligner1,plank,mark).
component(draw,winstr1,plank,mark).
neq(grasper1,grasper2).
grasper1 in {righthand, lefthand}.
grasper2 in {righthand, lefthand}.
aligner1 in {ruler}.  winstr1 in {pen}.
```

Again, a Simplis

An information we get from K-parser

If we assume the the knowledge that:

- lefthand and righthand are graspers,
- Ruler is an aligning artifact, and
- Pen is a writing instrument.

We get the resulting predicates:

```
component(g1,grasper1,plank,mark).
component(g2,grasper1,aligner1,mark).
component(g3,grasper2,aligner1,mark).
component(align,aligner1,plank,mark).
component(draw,winstr1,plank,mark).
neq(grasper1,grasper2).
grasper1 in {righthand, lefthand}.
grasper2 in {righthand, lefthand}.
aligner1 in {ruler}.  winstr1 in {pen}.
```

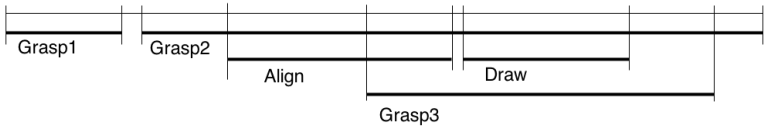
Optimizing Time Constraints

- Enumerate all possible before, subinterval, and startsbr4ov facts;
- Define pbefore, psubinterval and startsbr4ov interms of before, subinterval and startsbr4ov
- Use the given important facts about pbefore, psubinterval and startsbr4ov as constraints and minimize the number of before, subinterval, and startsbr4ov facts.

Optimization gives us..

```
pbefore(g1,g2).
pbefore(g1,align).
pbefore(g1,g3).
pbefore(g1,draw).
pbefore(align,draw).
psubinterval(align,g2).
psubinterval(g3,g2).
psubinterval(draw,g3).
pstartsb4ov(align,g3).
psubinterval(draw,g2).
```

Timeline representation



Minimized Constraints

```
before(g1,g2).
before(align,draw).
subinterval(g3,g2).
subinterval(align,g2).
subinterval(draw,g3).
startsb4ov(align,g3).
```

Summary

- Used state-of-the-art Visual recognition systems.
- Showed possible way to improve visual perception through knowledge and Common-sense Reasoning.
- Go beyond perception. Reason about objects, actions, subjects using knowledge.
- Possible way of activity recognition in limited-data scenario.

Future Work

- Currently working on Knowledge-base creation, that contains common day-to-day knowledge. Not factoid.
- Generalizing each one of the ideas mentioned. More complicated activity structures.
- Each question required new ASP program. Not feasible in real scenarios. How to generalize?

References I


Hand detection using multiple proposals.


E.E. Aksoy, A. Abramov, J. Dörr, K. Ning, B. Dellen, and F. Wörgötter.
Learning the semantics of object-action relations by observation.
The International Journal of Robotics Research, 30(10):1229–1249, 2011.


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 *Proceedings of the 2009 IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1932–1939, Miami, FL, 2009. IEEE.

References II

Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip H. S. Torr.
BING: Binarized normed gradients for objectness estimation at 300fps.
In *IEEE CVPR*, 2014.

Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie.
Behavior recognition via sparse spatio-temporal features.

In *Proceedings of the Second Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72, San Diego, CA, 2005. IEEE.

References III

Yangqing Jia.
Caffe: An open source convolutional architecture for fast feature embedding.
http://caffe.berkeleyvision.org/, 2013.

Jungseock Joo, Weixin Li, Francis F. Steen, and Song-Chun Zhu.
Visual persuasion: Inferring communicative intents of images.
In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 216–223, 2014.

I. Laptev.
On space-time interest points.
International Journal of Computer Vision, 64(2):107–123, 2005.

References IV


I. Oikonomidis, N. Kyriazis, and A. Argyros.
Efficient model-based 3D tracking of hand articulations using Kinect.
In *Proceedings of the 2011 British Machine Vision Conference*, pages 1–11, Dundee, UK, 2011. BMVA.


Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba.
Inferring the why in images.
CoRR, abs/1406.5472, 2014.


C. Teo, Y. Yang, H. Daume, C. Fermüller, and Y. Aloimonos.
Towards a Watson that sees: Language-guided action recognition for robots.
In *Proceedings of the 2012 IEEE International Conference on Robotics and Automation*, pages 374–381, Saint Paul, MN, 2012. IEEE.

References V

Dan Xie, Sinisa Todorovic, and Song-Chun Zhu.
Inferring "dark matter" and "dark energy" from videos.
In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 2224–2231, 2013.

Yezhou Yang, Cornelia Fermüller, and Yiannis Aloimonos.
Detection of manipulation action consequences (MAC).
In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2563–2570, Portland, OR, 2013. IEEE.