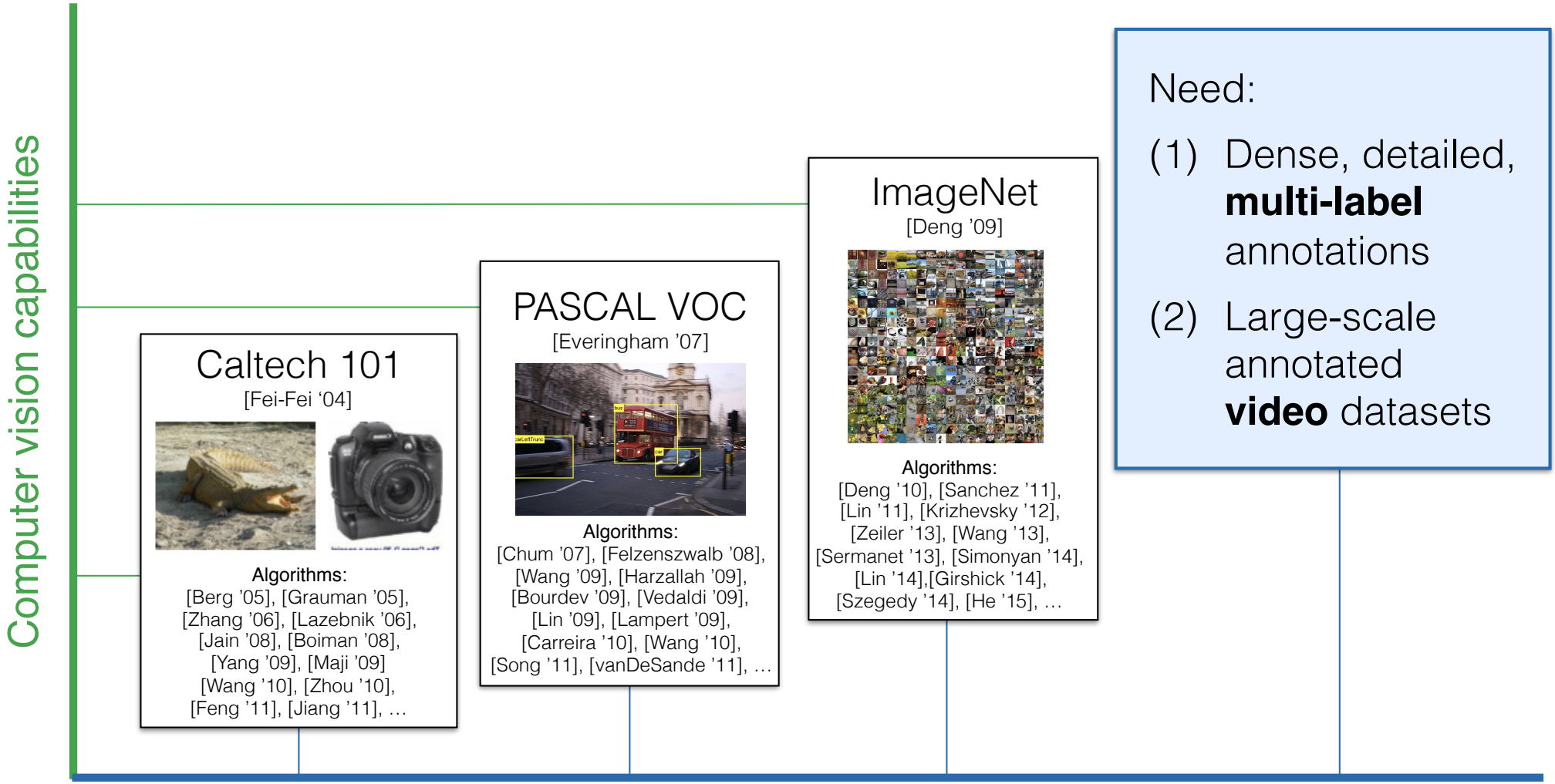
Much Ado About Time: Exhaustive Annotation of Temporal Data

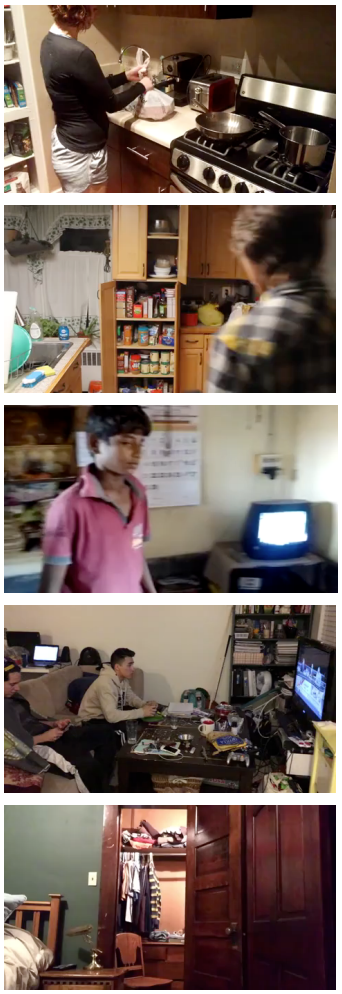
Gunnar A. Sigurdsson, Olga Russakovsky,
Ali Farhadi, Ivan Laptev, Abhinav Gupta

Carnegie Mellon University ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE UNIVERSITY of WASHINGTON Inria

Datasets drive computer vision progress



Multi-label video annotation

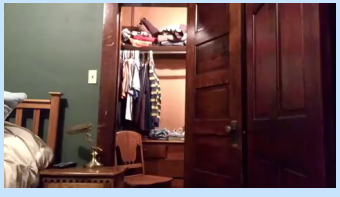


10,000 videos

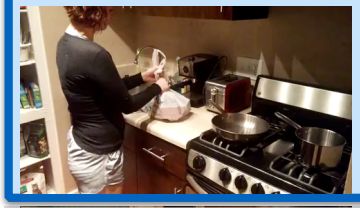
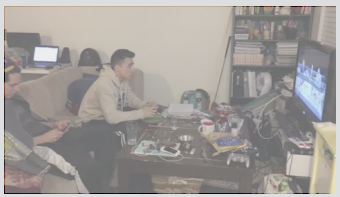
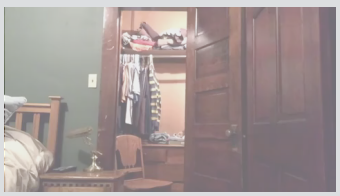
opens book	puts book on shelf	walks	turns on stove	eats	sits down	sneezes
-	-	-	+	-	-	-
-	+	+	-	-	-	+
-	-	+	-	-	+	-
-	-	-	-	+	-	-
+	-	+	-	-	+	-

100-200 labels

Multi-label video annotation

	opens book	puts book on shelf	walks	turns on stove	eats	sits down	sneezes
	?	-	-	+	-	-	-
	?	+	+	-	-	-	+
	?	-	+	-	-	+	-
	?	-	-	-	+	-	-
	?	-	+	-	-	+	-

Multi-label video annotation

	opens book	puts book on shelf	walks	turns on stove	eats	sits down	sneezes
	?	?	?	?	?	?	?
	-	+	+	-	-	-	+
	-	-	+	-	-	+	-
	-	-	-	-	+	-	-
	+	-	+	-	-	+	-

Which interface is better?

One-label

☐ Opens book

Repeat N times for N labels

vs

All-labels

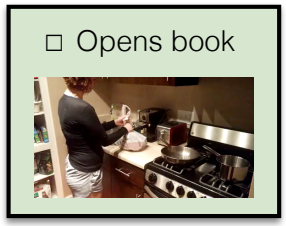
☐ Opens book
☐ Puts book on shelf
☐ Walks
☐ Turns on stove
☐ Eats
☐ Sits down
…

Expect better annotation ***accuracy***

Expect better annotation ***time***

Which interface is better?

One-label



Repeat N times for N labels

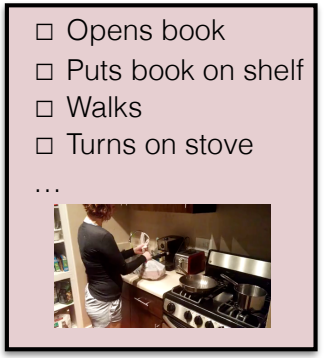
Data: 140 videos, each ~30 secs long

Labels: 52 human actions

Charades dataset of [Sigurdsson ECCV 2016]

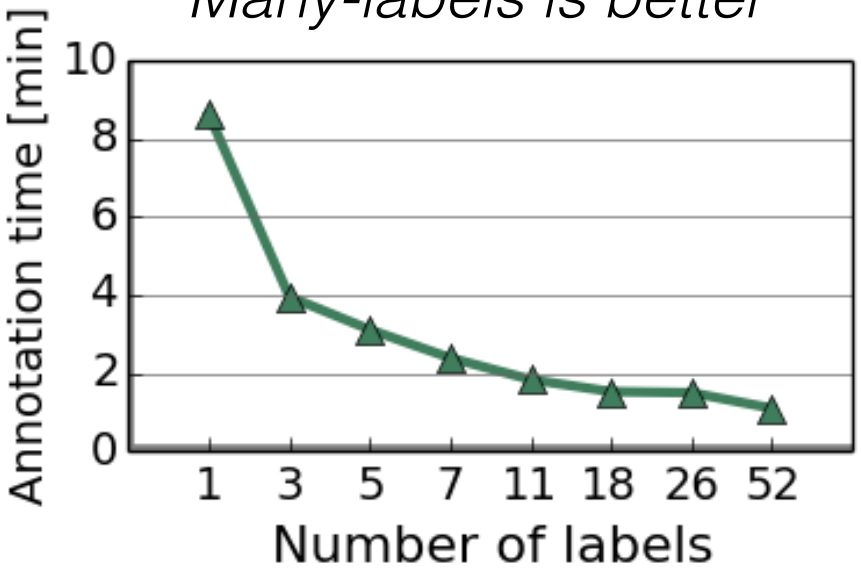
Experiment on Amazon Mechanical Turk

All-labels



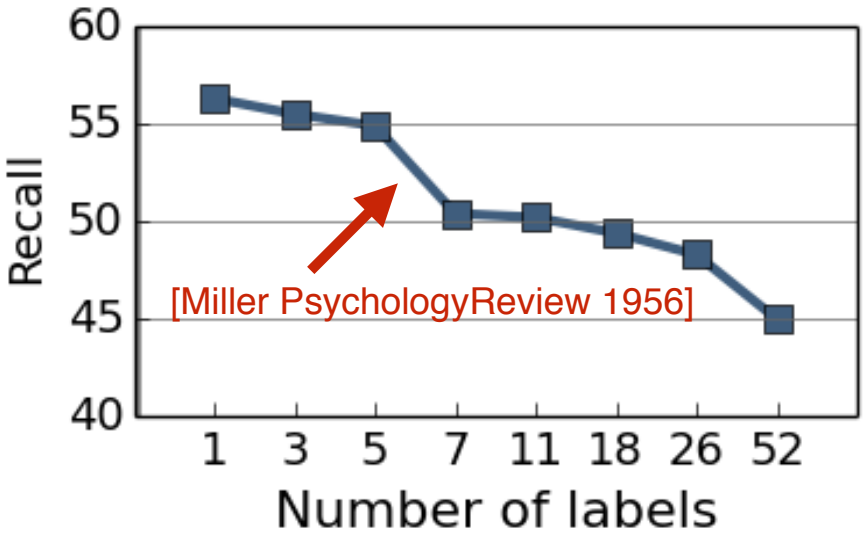
Time

Many-labels is better

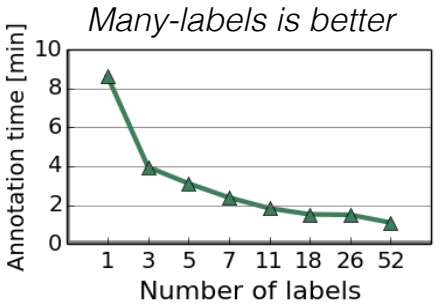


Accuracy

Few-labels is better



Improving annotation time

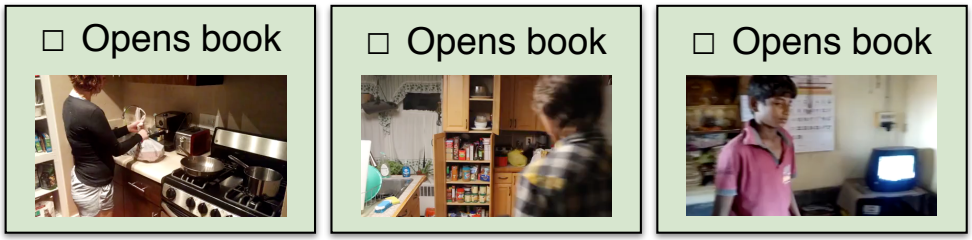


Consistency in the few-labels setting

Ask same worker about the same actions for multiple videos

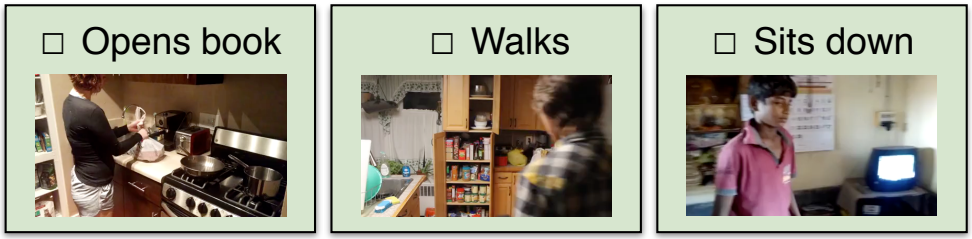
=> 13.6% reduction in annotation time

Worker 1:



vs

Worker 1:

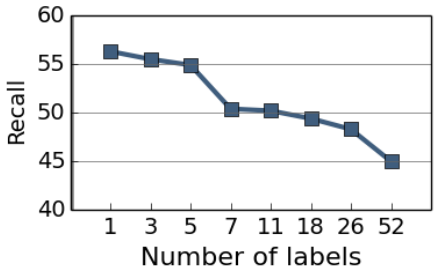


Play video at 2x speed [Lasecki UIST 2014]

Semantic hierarchy of labels [Deng CHI 2014]

Improving recall

Few-labels is better



Many-labels

- ☐ Opens book
- ☐ Puts book on shelf
- ☐ Walks
- ☐ Turns on stove
- ☐ Eats
- ☐ Sits down
- ☐ Sneezes
- ☐ Picks up a cup
- ☐ Holds a dish

…

Video summary

Request a 20-word description of the video

=> no effect on recall, 40% slower

Forced response

Request a yes/no response for every label

=> actually drops recall! (annoys workers?)

Consensus annotation

Rely on multiple rounds of annotation with different workers

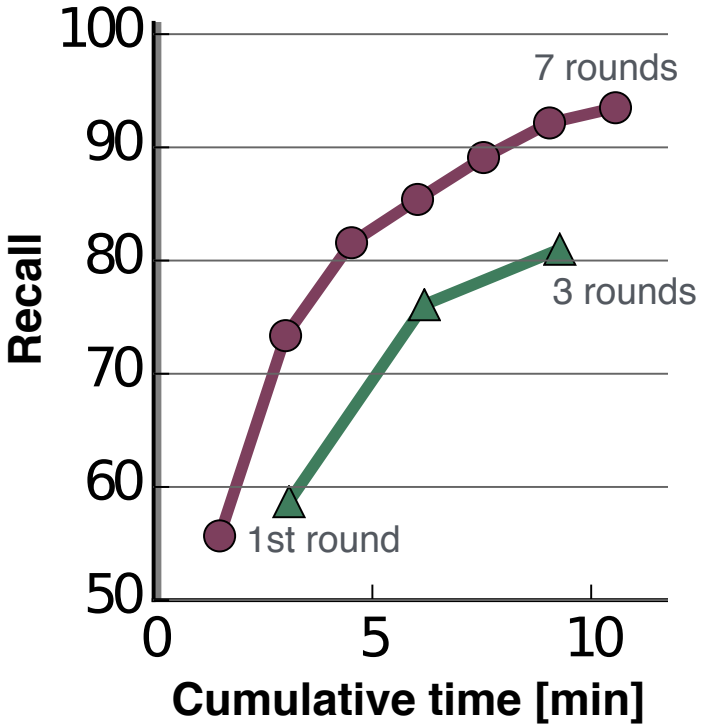
=> recall improves from 58.0% to 83.3% with 3 rounds

[Krishna CHI 2016]

Bringing it all together

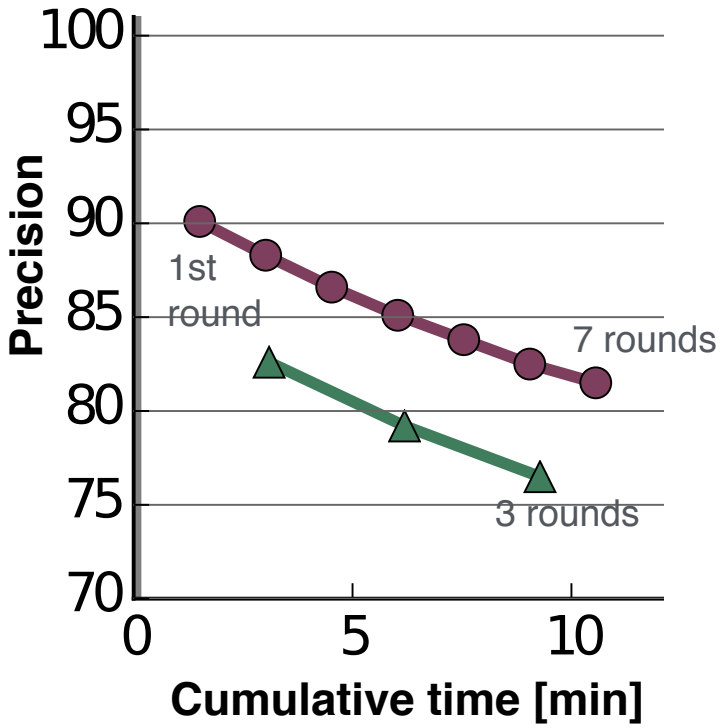
Data: 1,815 videos, each ~30 secs long, 2x speed
Labels: 157 human actions, organized into a hierarchy with 52 high-level actions
Charades dataset of [Sigurdsson ECCV 2016]

Experiments on Amazon Mechanical Turk
Label is positive if >= 1 worker marks it as positive



Many-label interface (26)

Few-label interface (5)



Conclusions

- **Quantitative** analysis of multi-label video annotation
- **Many-labels** interface is better than the **few-labels** interface
- Annotated of **157 human actions** on **9,848 videos** (incl. temporal extent)

Download dataset at http://allenai.org/plato/charades

Actions

Holding a dish
Holding a cup/glass/bottle of something
Walking through a doorway
Someone is standing up from somewhere

Video (3x speed)