Gunnar A. Sigurdsson, Olga Russakovsky, and Abhinav Gupta

What Actions are Needed for Understanding Human Actions in Videos?

THE ROBOTICS INSTITUTE

Carnegie Mellon University

Activity Data

What is the right data? How to use the data?

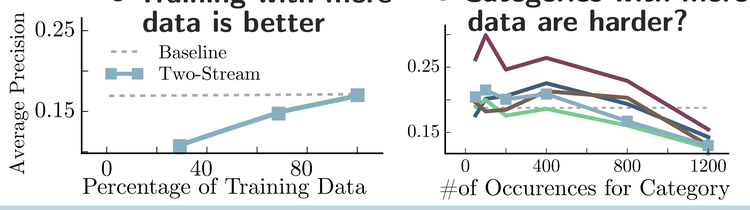
Human Interpretation

- Multiple subjects annotate the actions and are compared

Takeaway: Action boundaries are fluid

Machine Interpretation

- Training with more data is better
- Categories with more data are harder?



Takeaways: 1. No balanced data–Classes are different
2. Special attention: small/large categories

Evaluation Setup

- Collect various attributes for multiple datasets
- Train action classification/localization baselines
- Evaluate video/frame mAP on:

Algorithms:

Two-Stream (Simonyan/Zisserman)
IDT (Wang/Schmid)
CNN+LSTM (Ng et al.)
ActionVLAD (Girhar et al.)
TFields (Sigurdsson et al.)

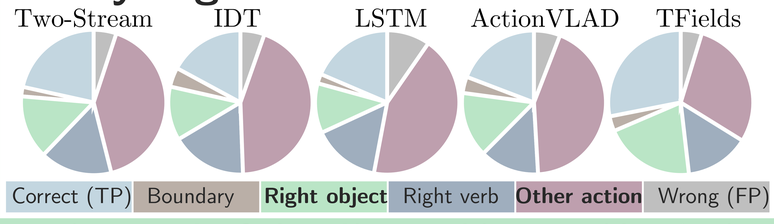
Datasets:

Charades (Sigurdsson et al.)
ActivityNet (Heilbron et al.)
MultiTHUMOS (Yeung et al.)
THUMOS'14 (Gorban et al.)

Improving Algorithms

How can we improve state-of-the-art algorithms?

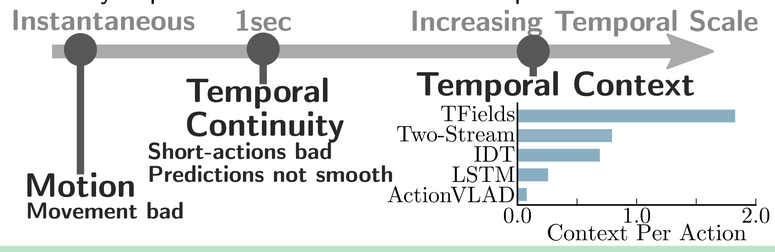
Analyzing Detections



Takeaway: Confusion–Classes with same object/verb

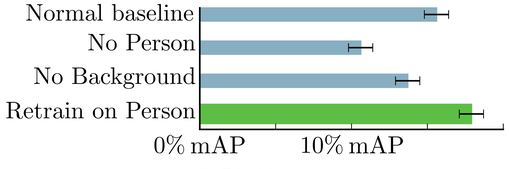
Temporal Reasoning

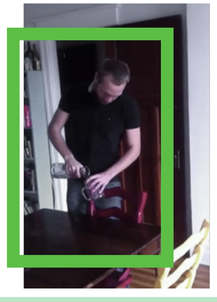
- Analyze performance with various temporal attributes



Takeaway: Algorithms can benefit from temporal reasoning on all temporal scales

Person-based Reasoning





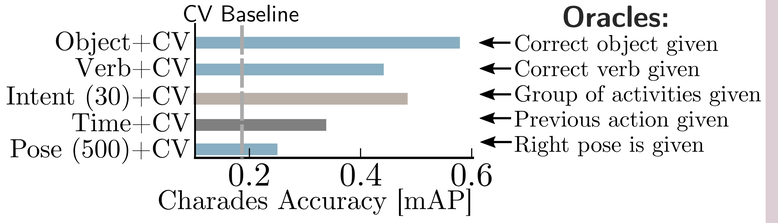
- Training on RCNN person boxes significantly increases performance.

Takeaway: Teach algorithms about people.

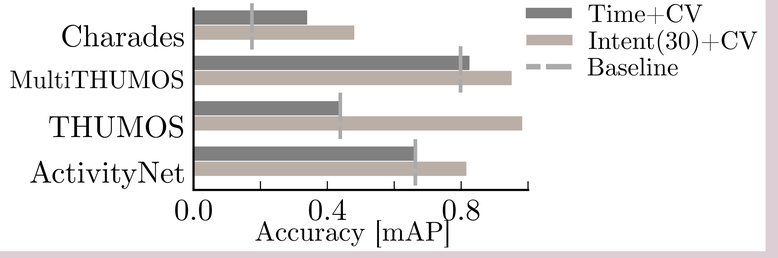
Promising Directions

What cues are likely to yield big gains?

- Evaluate different types of perfect information on datasets



Takeaway: Object+Time would yield substantial gains



Takeaways: 1. Some datasets suitable for time models
2. All benefit from global reasoning

Available Resources

ECCV'16
CVPR'17
(Challenge)

- Charades Dataset has many diverse attributes
- Useful benchmarking and analyzing algorithms
- Attributes/Code to diagnose any algorithm:

github.com/gsig/actions-for-actions