What Actions are Needed for Understanding Human Actions in Videos?

Gunnar A. Sigurdsson, Olga Russakovsky, and Abhinav Gupta

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

Human Interpretation
- Multiple subjects annotate the actions and are compared
  - More confusion about verb than object
  - 72.5% IOU Agreement (1.4s end, 0.9s start)

Takeaway: Action boundaries are fluid
- Training with more data is better
- Categories with more data are harder?

Machine Interpretation
- Training with different temporal attributes
  - Instantaneous
  - 1sec
  - Increasing Temporal Scale
  - Temporal Continuity
  - Short actions bad
  - Motion
  - Predictions not smooth
  - Movement bad

Takeaway: Confusion–Classes with same object/verb
- Temporal Reasoning
- Takeaway: Algorithms can benefit from temporal reasoning on all temporal scales

Evaluation Setup
- Collect various attributes for multiple datasets
- Train action classification/localization baselines
- Evaluate video/frame mAP on:
  - Normal baseline
  - No Person
  - No Background
  - Retrain on Person

Promising Directions
What cues are likely to yield big gains?
- Evaluate different types of perfect information on datasets
  - CV Baseline
  - Verbal
  - Intent (30)
  - Time
  - Pose (500)

Takeaway: Object+Time would yield substantial gains
- Time+CV
- Intent (30)+CV
- Baseline

Available Resources
- Charades Dataset has many diverse attributes
- Useful benchmarking and analyzing algorithms
- Attributes/Code to diagnose any algorithm:
github.com/gsig/actions-for-actions

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

Analyzing Detections
- Two-Stream
- IDT
- LSTM
- ActionVLAD
- TFields

Takeaway: Confusion–Classes with same object/verb
- Temporal Reasoning
- Takeaway: Algorithms can benefit from temporal reasoning on all temporal scales

Person-based Reasoning
- Normal baseline
- No Person
- No Background
- Retrain on Person

Takeaway: Teach algorithms about people.

ECCV’16 CVPR’17 (Challenge)