Moments: A Large-Scale Database for Video Understanding

Mathew Monfort¹, Tom Yan¹, Bolei Zhou¹, Dan Gutfreund², Quanfu Fan², Carl Vondrick¹, Aude Oliva¹

¹MIT CSAIL, ²IBM

Overview

- 2.13 million temporally localized 3-second videos
- 700,000 positive labels
- 1.64 million negative labels
- 353 “moment” classes
- Multi-label annotation
- High-level conceptual moments
- Scene transformation/changes
- Breadth of moment examples

Data Distribution

- Balanced across classes
- More than 1,000 videos per action
- Human, animal, object and natural phenomena agents
- Visual and auditory annotation

Moment Recognition Baselines

- 3-stream baseline model
  - Spatial
  - Temporal
  - Auditory
  - Fusion via multi-class linear SVM
- Human (H) vs non-Human (NH) agent comparison
- Learned class similarity
- Spatial attention analysis

Scene, Object and Action Correlation

- Networks trained on both ImageNet (objects) and Places (scenes)
- Bayesian inference
- 10-fold cross validation
- Importance of audio stream