Hollywood in Homes: Crowdsourcing Data Collection For Activity Understanding

Introduction & Motivation

What is **Hollywood in Homes**?

- Crowdsourcing data collection (On Amazon Mechanical Turk)
- New approach to collect data for Computer Vision
- Created new dataset called Charades v1.0

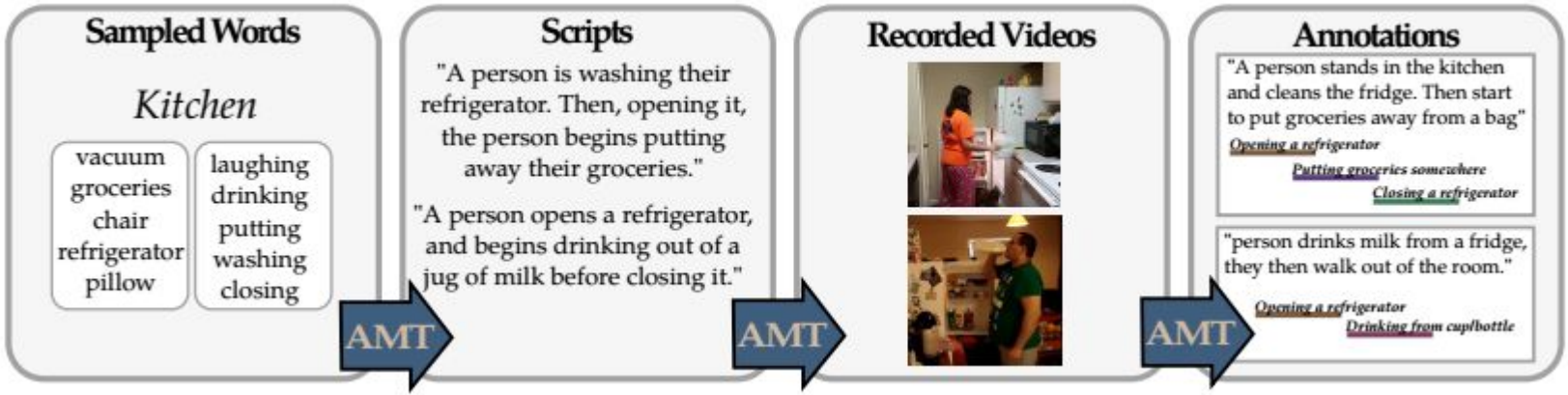
The Charades Dataset

YouTube

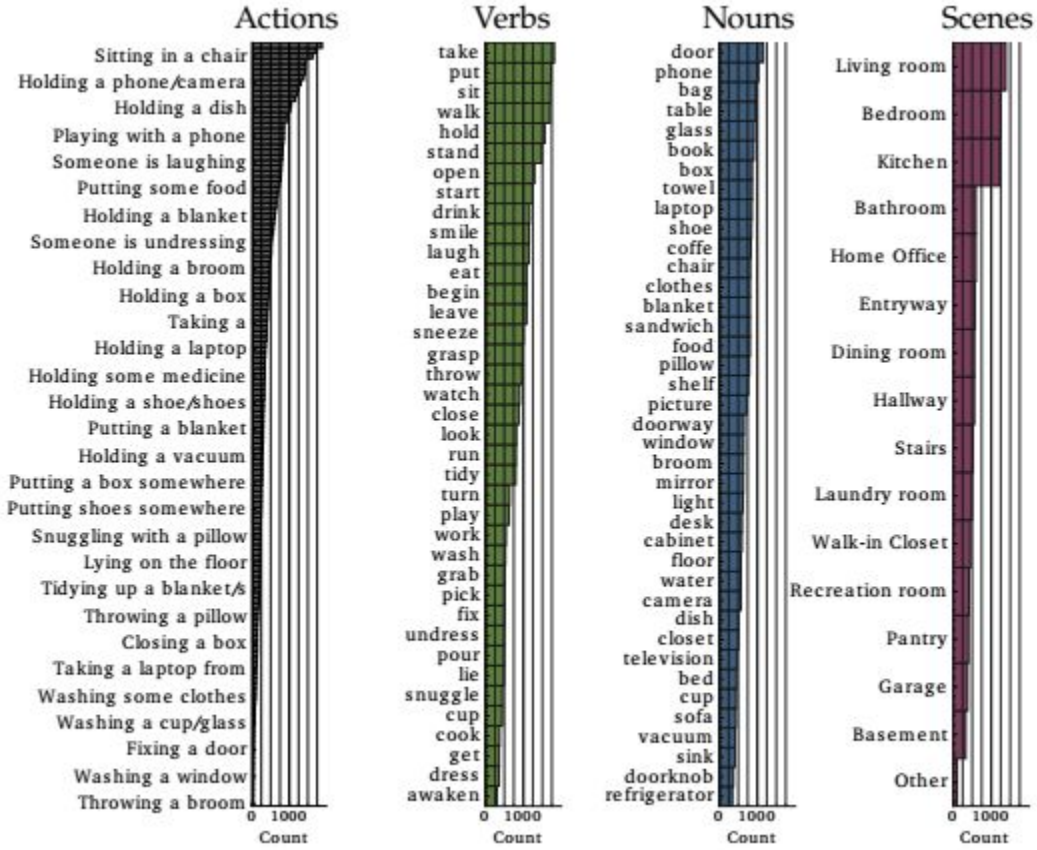
Introduction & Motivation

Three steps of filming process:

1. Script writing
2. Video direction and acting based on scripts
3. Video verification



Charades v1.0



Charades v1.0

Annotated Actions: (gray if not active)

Sitting in a chair

Holding a pillow

Snuggling with a pillow

Video 32 of 50: (3x Speed)

Annotated Objects:

Chair, Game, Pillow

Script:

A person is sitting on a chair with a pillow watching someone play a game.

Introduction & Motivation

Why do we need this new data set?

- Current data set is often biased toward static scenes & objects in Internet images
- Need to learn different states of objects, how activities affect change of object states
- Need datasets about boring activities in our life, which is very limited on the Internet

Charades vs. other video datasets

Table 1. Comparison of Charades with other video datasets.

	Actions per video	Classes	Labelled instances	Total videos	Origin	Type	Temporal localization
Charades v1.0	6.8	157	67K	10K	267 Homes	Daily Activities	Yes
ActivityNet [3]	1.4	203	39K	28K	YouTube	Human Activities	Yes
UCF101 [8]	1	101	13K	13K	YouTube	Sports	No
HMDB51 [7]	1	51	7K	7K	YouTube/Movies	Movies	No
THUMOS'15 [5]	1-2	101	21K+	24K	YouTube	Sports	Yes
Sports 1M [6]	1	487	1.1M	1.1M	YouTube	Sports	No
MPII-Cooking [14]	46	78	13K	273	30 In-house actors	Cooking	Yes
ADL [25]	22	32	436	20	20 Volunteers	Ego-centric	Yes
MPII-MD [11]	Captions	Captions	68K	94	Movies	Movies	No

Application

- Run several state of the art algorithms on Charades to provide baselines for recognizing human activities in realistic home environments
- Train/Test set
  - No worker crossover
  - Similar distribution of categories (min. of 6 test and 25 training videos per category)
  - Test set not dominated by a single worker
  - 7,985 training and 1,863 test videos, with 49,809 and 16,691 annotated action intervals respectively

Application: Action Classification

- Action Classification
  - Given a video, identify whether it contains any of the 157 action classes
  - Classification performance evaluated with mean average precision (mAP)
- The Classification Baselines (mAP)
  - C3D Features 10.9%
  - Static CNN Features 11.3%
  - Balanced Two-stream Networks 11.9%
  - Two-stream Networks 14.3%
  - Improved Dense Trajectory (IDT) Features 17.2%
  - Combined (late fusion) 18.6%

Static CNN Features -- 11.3%

- Authors experimented with VGG-16 and AlexNet
- Extracted fc_6 features over 30 equidistant frames
- Features were averaged across frames, L2-normalized, then classified with a one-versus-rest linear SVM

C3D (2015) -- 10.9%

- 3D Convolutional Network
  - Captures complex hierarchies of spatio-temporal patterns, 1vR linear SVM
- Volumetric Convolution, 4D Tensor with added time dimension
-

```
module = nn.VolumetricConvolution(nInputPlane, nOutputPlane, kT, kW, kH [, dT, dW, dH, padT, padW, padH])
```

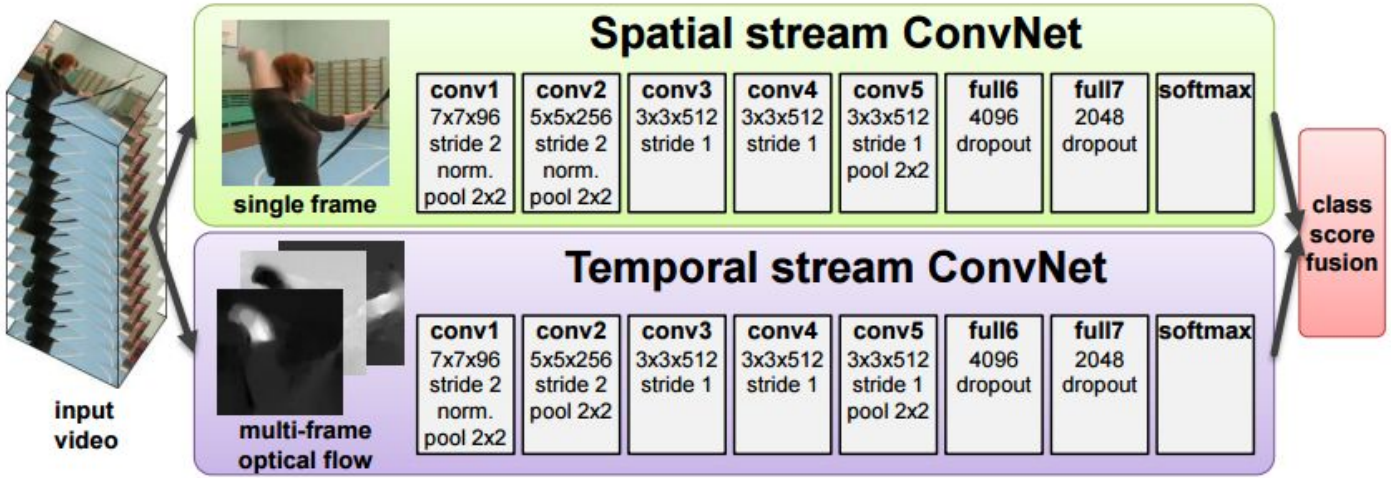
Kernel size of convolution in time

Step in time

Time zero-padding

Two-stream Networks (2014 x2) -- 11.9%, 14.3%

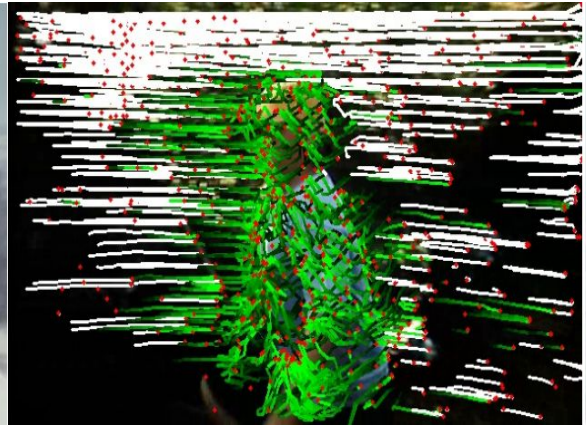
- Parallel spatial and temporal networks (VGG-16)
- For balanced, each minibatch of 256 had at least 50 unique action classes
- Spatial uses still image action recognition, Temporal uses optical flow



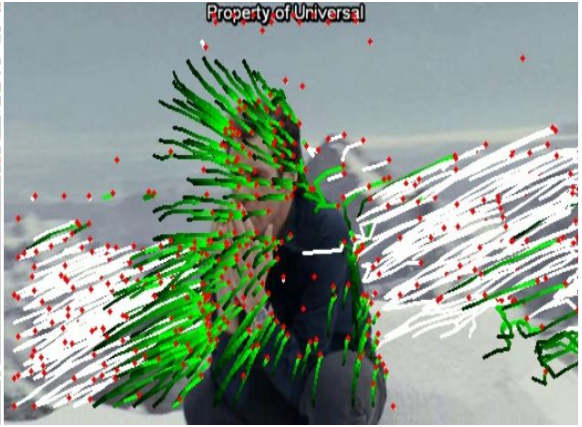
Improved Dense Trajectory (2011) -- 17.2%

Hist. of Oriented Gradients + Hist. of Optical Flow + Motion Boundary Histograms → Principal Comp. Analysis → GMM → Fisher Vectors → 1vR Linear SVM

2 overlaid frames

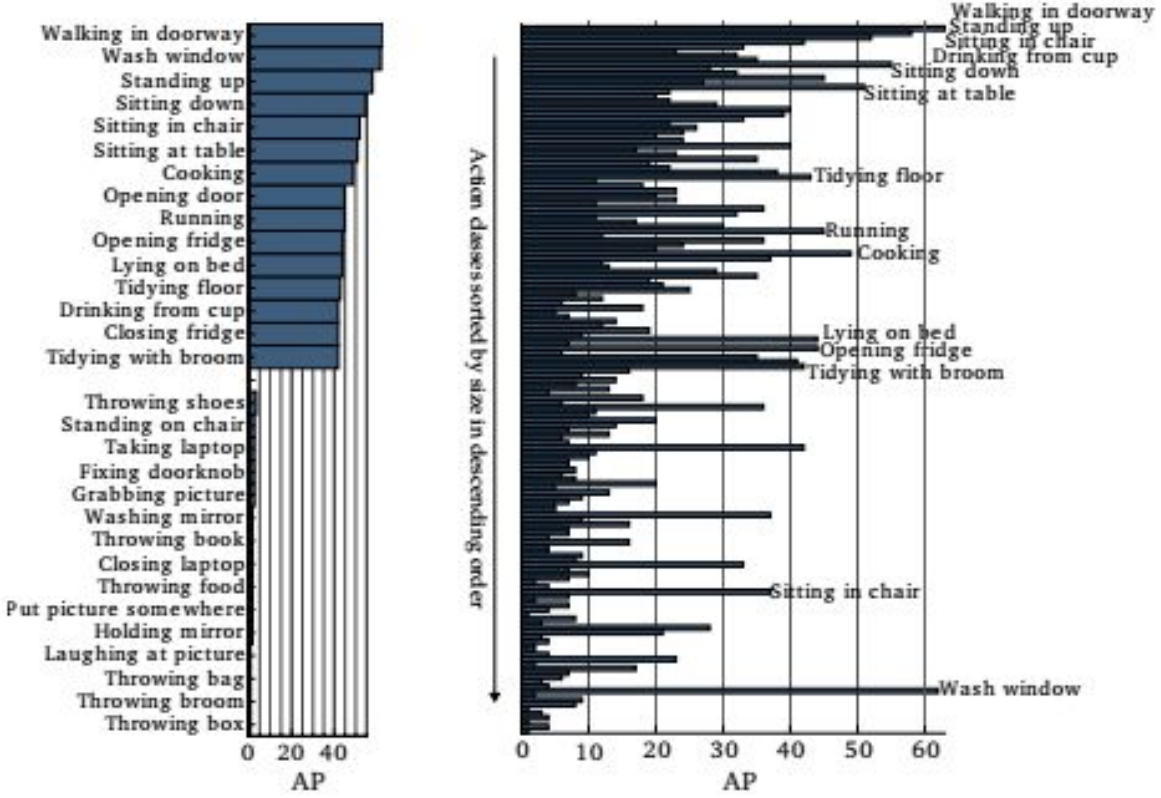


Optical Flow between Frames

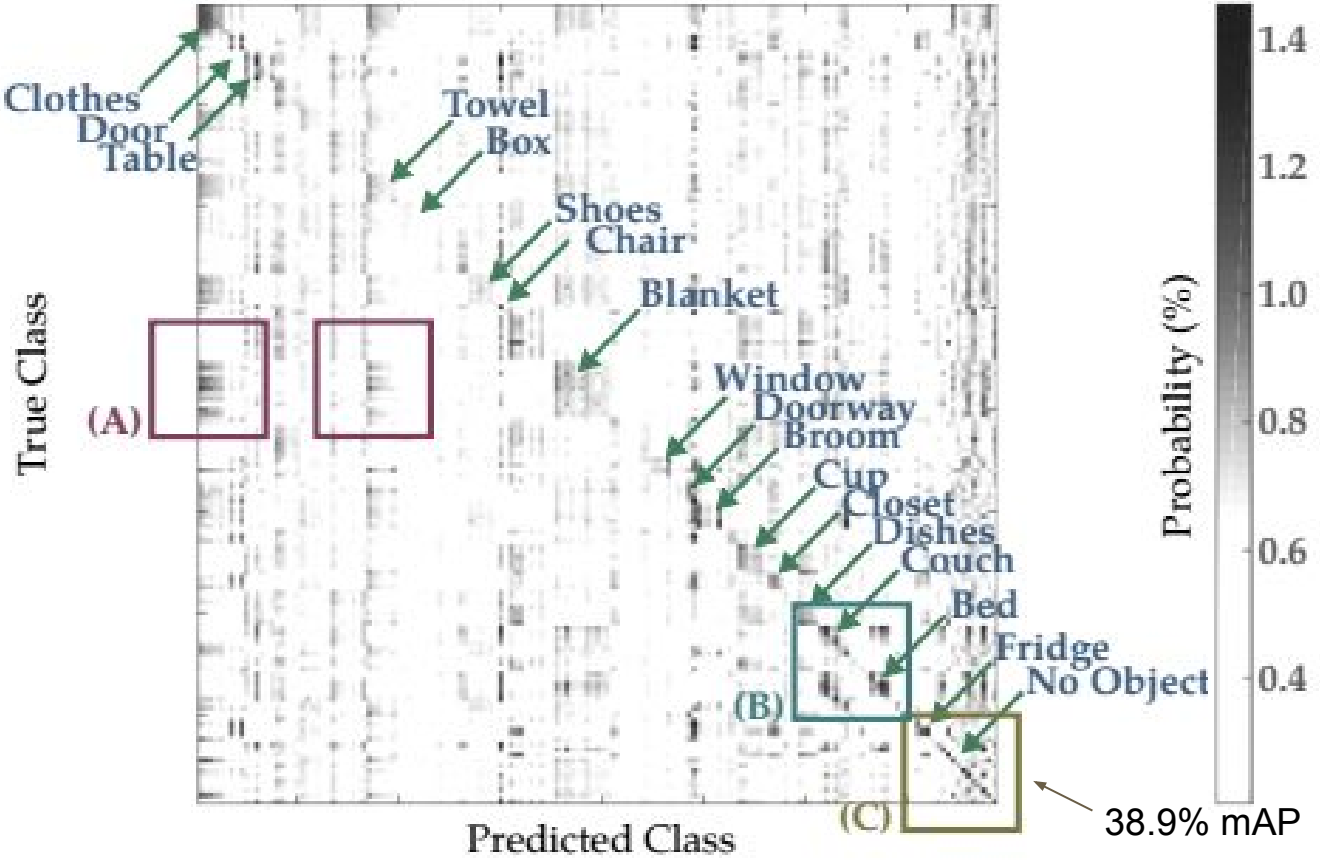


Trajectories

Application: Combined Baseline -- 18.6%



Application: Combined Confusion Matrix



Application: Sentence Prediction

- Sentence Prediction: free-form sentences that describe the video
- Script uses 1 sentence ground truth (GT), Description ~2.4 sentences GT
- Best performance is S2VT-- Sequence to Sequence: Video to Text (2015)

	Script					*Description*				
	RW	Random	NN	S2VT	Human	RW	Random	NN	S2VT	Human
CIDEr	0.03	0.08	0.11	0.17	0.51	0.04	0.05	0.07	0.14	0.53
$BLEU_4$	0.00	0.03	0.03	0.06	0.10	0.00	0.04	0.05	0.11	0.20
$BLEU_3$	0.01	0.07	0.07	0.12	0.16	0.02	0.09	0.10	0.18	0.29
$BLEU_2$	0.09	0.15	0.15	0.21	0.27	0.09	0.20	0.21	0.30	0.43
$BLEU_1$	0.37	0.29	0.29	0.36	0.43	0.38	0.40	0.40	0.49	0.62
$ROUGE_L$	0.21	0.24	0.25	0.31	0.35	0.22	0.27	0.28	0.35	0.44
METEOR	0.10	0.11	0.12	0.13	0.20	0.11	0.13	0.14	0.16	0.24

Conclusion & Future Works

- Proposed a new approach for building datasets
  - Crowdsourcing not only labeling, but also data gathering
  - Recyclable framework
- Built a large-scale dataset with diversity and unique realism
  - Realistic object-action relationships (46 objects, 30 actions, 15 indoor scenes)
- Provided baselines that enables benchmarking for future algorithms
- Inspire exploration of novel domains and development of novel computer vision techniques in object-action relationships