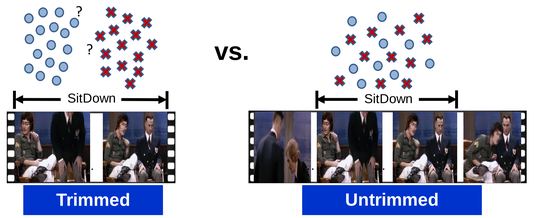
Active Learning of an Action Detector from Untrimmed Videos

Sunil Bandla and Kristen Grauman
Department of Computer Science, University of Texas at Austin

Motivation

- Realistic unlabeled videos are "untrimmed" to temporal regions of interest, and each video contains multiple actions.
- This yields unlabeled feature distribution where useful and redundant candidates are hard to distinguish for active learning.



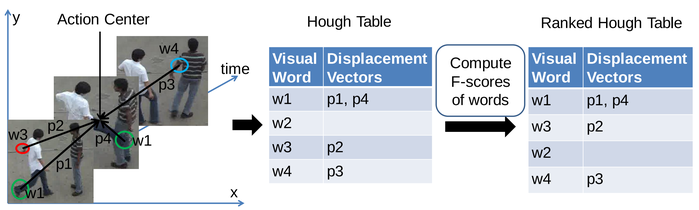
Main Idea

- We introduce a detection-based active learning approach to select videos for annotation, while accounting for their untrimmed nature.
- Voting-based detector is robust to partial evidence and supports fast incremental updates during active learning.
- Learn accurate action recognition models with fewer annotations.

Hough-based Action Detector

Building the Detector

- Extract HoG/HoF features at STIPs detected in training videos and build Hough tables, and sort words by discriminative power.

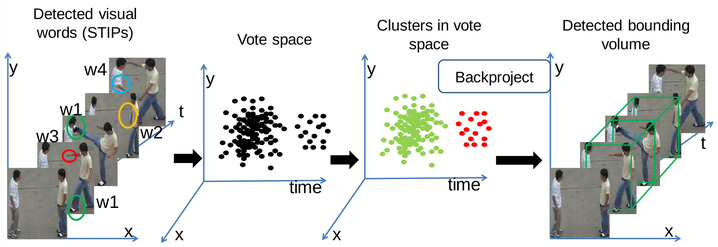


Visual Word	Displacement Vectors
w1	p1, p4
w2	
w3	p2
w4	p3

Visual Word	Displacement Vectors
w1	p1, p4
w3	p2
w2	
w4	p3

Applying the Detector to a Novel Video

- Use the Hough table entries to vote on the probable action centers
- Reduces number of candidate intervals per video for active selection



[Similar to detectors of Willems et al., BMVC 2009; Yao et al., CPVR 2010]

Active Selection of Untrimmed Videos

We seek the unlabeled video that, if used to augment the action detector, will more confidently localize actions in *all* unlabeled videos.

$$v^* = \underset{v \in \mathcal{U}}{\operatorname{argmax}} \quad \max_{l \in \mathcal{L}} S(\mathcal{T} \cup v^l),$$

where $\mathcal{T}$ is the training set, $\mathcal{L}$= {+1,−1} is the set of possible labels, and v^l denotes that the unlabeled video v has been given label l.

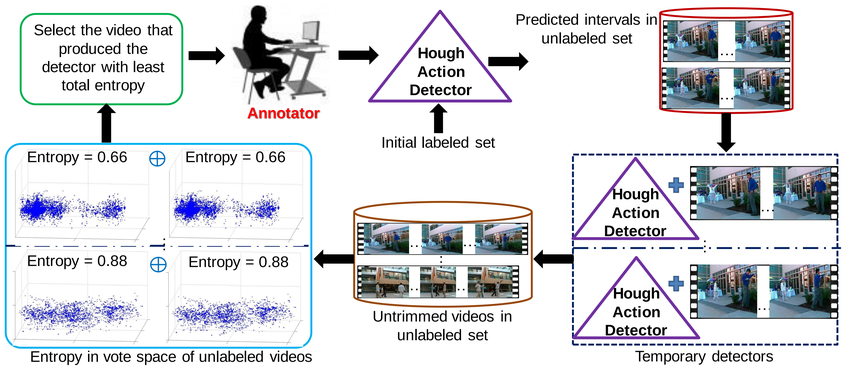
- Treating the unlabeled video v as positive, we score the value of probable action intervals in the video to the current detector D:

$$S(\mathcal{T} \cup v^+) = \max_{k=1,\ldots,K} \textsc{Value}\big(D(\mathcal{T} \cup \hat{v}_k^+)\big)$$

- Treating v as negative,

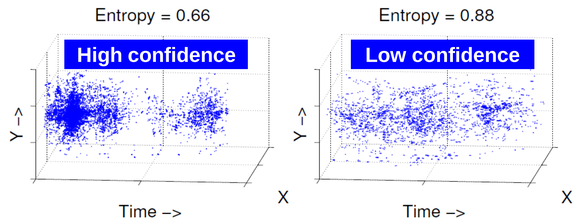
$$S(\mathcal{T} \cup v^-) = \textsc{Value}\big(D(\mathcal{T} \cup v^-)\big)$$

where VALUE is our novel entropy-based detector confidence defined below.



Estimating Detector Confidence with Space-Time Entropy

- Quantize unlabeled video's 3D vote space and compute normalized entropy
- A vote space with good cluster(s) indicates consensus on the location(s) of the action



- Using this entropy-based uncertainty metric H, we define the confidence of a detector $D(\mathcal{T})$ in localizing actions on the entire unlabeled set $\mathcal{U}$:

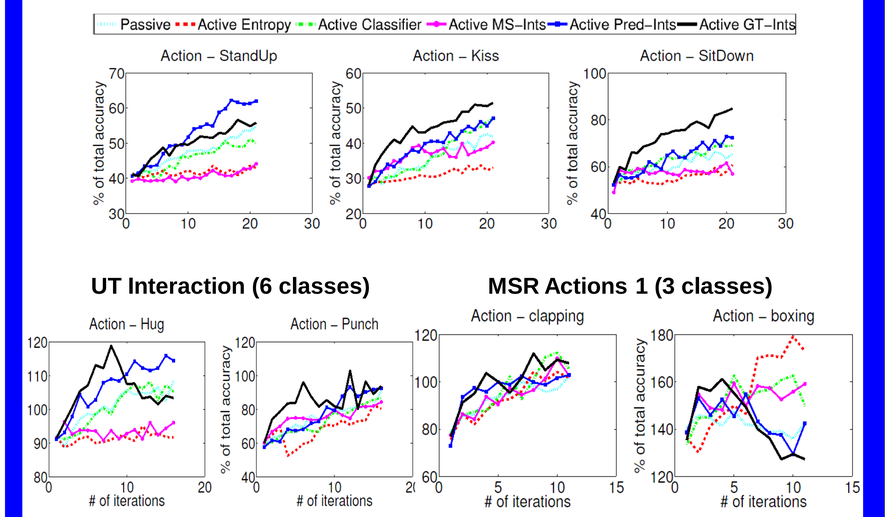
$$\textsc{Value}\big(D(\mathcal{T})\big) = \frac{1}{|\mathcal{U}|} \sum_{v \in \mathcal{U}} \big(1 - H\big(V_v, D(\mathcal{T})\big)\big)$$

Annotations

Our interface that annotators use to label action intervals in the actively requested videos.

Available on the project webpage.

Results



- **Passive < Active**: Annotation effort saved by intelligent label requests.
- **Active Classifier < Ours**: Accounting for *untrimmed* nature of video is critical.
- **Active Entropy < Ours**: Simply estimating *individual* video uncertainty is insufficient.
- **Active GT-Ints > Active Pred-Ints**: Room for improvement in interval estimates.

Train Set	HandShake	Hug	Kick	Point	Punch	Push
Initial L ex only	0.1981	0.3029	0.1466	0.0107	0.1094	0.2022
After 15 rounds active	0.2574	**0.3804**	0.2175	0.0164	0.1758	0.2689
Full train set (42 ex)	**0.2708**	0.3324	**0.3218**	**0.0478**	**0.1897**	**0.3058**

UT Interaction

Train Set	Clapping	Waving	Boxing
Initial L ex only	0.2288	0.2318	**0.1135**
After 10 rounds active	**0.3379**	**0.3134**	0.1043
Full train set (27 ex)	0.3132	0.2582	0.0819

MSR Actions 1

Our active method achieves good accuracy using much less annotations.