# Watching Unlabeled Video Helps Learn New Human Actions from Very Few Labeled Snapshots

Chao-Yeh Chen and Kristen Grauman
The University of Texas at Austin

## Problem

Goal: learn actions from **static images**

Problems of static snapshots:
- May have only few training examples for some actions.
- Often limited to “canonical” instances of the action.

## Our idea

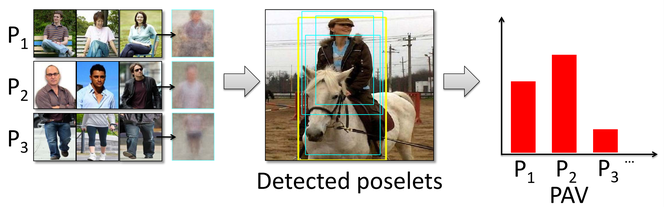
Expand snapshots by pose dynamics learned from videos


**Let the system:**
- Watch videos to learn how human poses change over time.
- Infer nearby poses to expand the sparse training snapshots.

## Related Work

- Learn actions with discriminative pose and appearance features.
  e.g. [Maji et al. 2011, Yang et al. 2010, Yao et al. 2010, Delaitre et al. 2011]
- Expand training data by mirroring images and videos.
  e.g. [Papageorgiou et al. 2000, Wang et al. 2009]
- Synthesize images for action recognition and pose estimation.
  e.g. [Matikainen et al. 2011, Shakhnarovich et al. 2003, Grauman et al. 2003, Shotton et al. 2011,]
- Ours: expanding the training set for “free” via pose dynamics learned from unlabeled data.

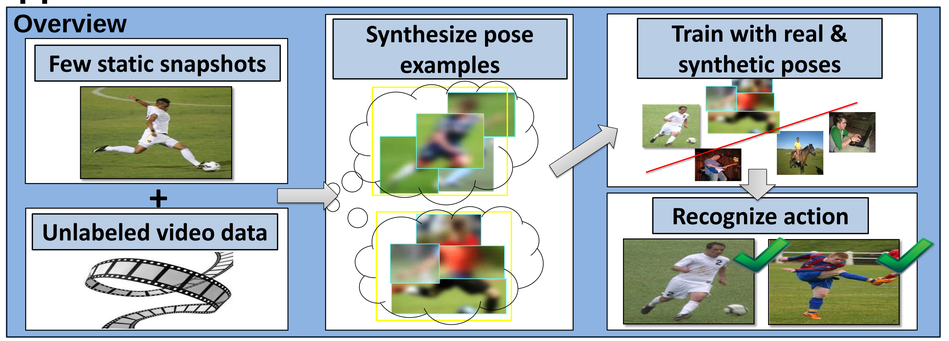
## Representing body pose



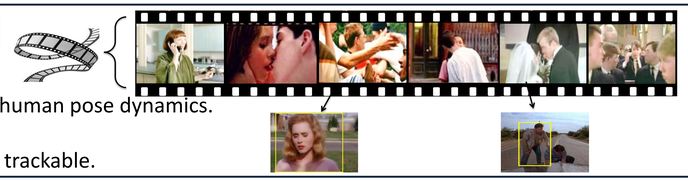

- Poselet activation vector (PAV) by [Maji et al. 2011]
- Each poselet captures part of the pose from a given viewpoint
- Robust to occlusion and cluttered background

## Approach

### Overview

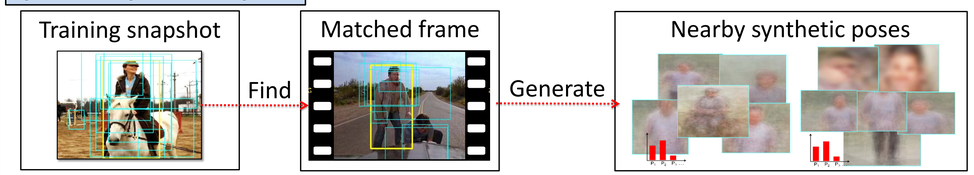



### Unlabeled video data

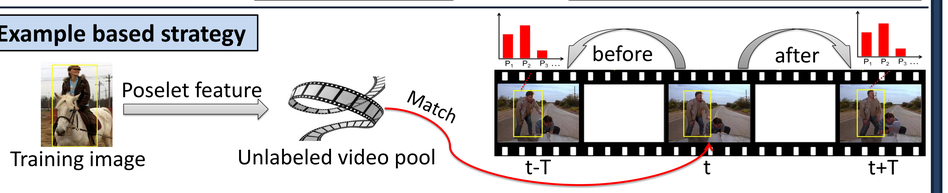


Assumptions:
- Videos cover the space of human pose dynamics.
- No action labels are given.
- People are detectable and trackable.

### Synthesize pose examples:



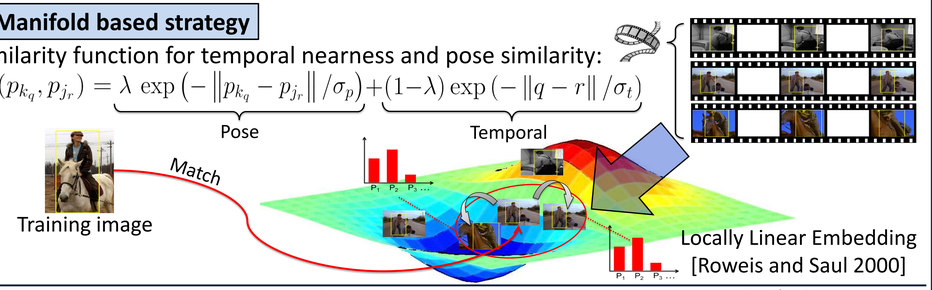

### 1) Example based strategy




### 2) Manifold based strategy

Similarity function for temporal nearness and pose similarity:

$$A(p_{k_q}, p_{j_r}) = \underbrace{\lambda \exp\left(-\left\|p_{k_q} - p_{j_r}\right\| / \sigma_p\right)}_{\text{Pose}} + \underbrace{(1-\lambda) \exp\left(-\left\|q - r\right\| / \sigma_t\right)}_{\text{Temporal}}$$

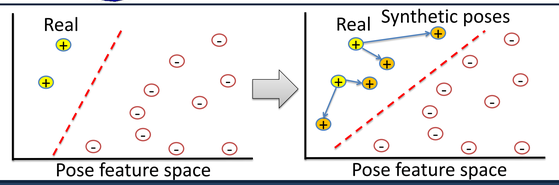



### Train with real & synthetic poses

Domain adaptation by [Daume III 2007]

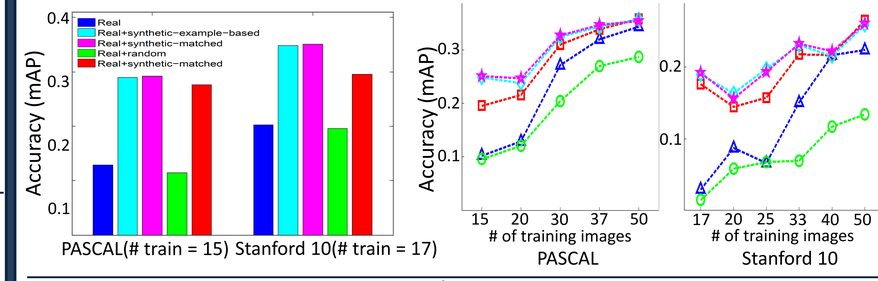
Static snapshots (target): $p^{i'} = [p^i, 0, p^i]$

Frames from video (source): $p^{v'} = [p^v, p^v, 0]$

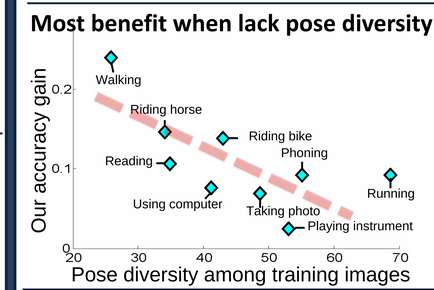



## Results

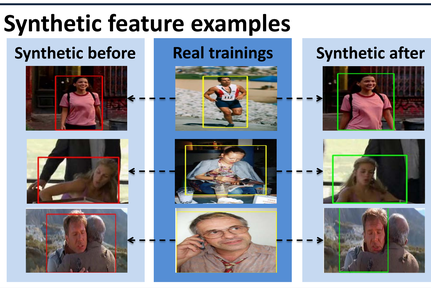

### Recognizing activity in images



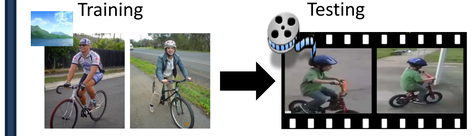

### Most benefit when lack pose diversity



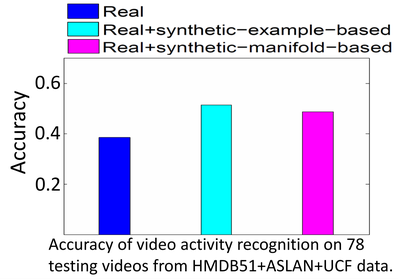

### Synthetic feature examples




### Recognizing activity in videos




- Substantial gains when recognizing actions in video, since our method infers intermediate poses not covered in original snapshots.




## Conclusions

- Augment training data without additional labeling cost by leveraging unlabeled video.
- Simple but effective exemplar/manifold extrapolation strategies.
- Significant advantage when labeled training examples are sparse.
- Domain adaptation connects real and generated examples.