Activities as Time Series of Human Postures

William Brendel and Sinisa Todorovic

11th European Conference on Computer Vision
ECCV 2010 [5-11 September, 2010]
Hersonissos, Heraklion, Crete, Greece

MOTIVATION

Activity recognition should be:

a) Robust: dynamic backgrounds, occlusion, clutter

b) Efficient: fast processing

c) Scalable: large number of classes, large datasets

PROBLEM STATEMENT

input videos sequence of poses

Given a set of videos with class labels,

Extract a dictionary of:

- Snapshots of human postures,
- Co-occurring objects (e.g., tools).

Represent a video as

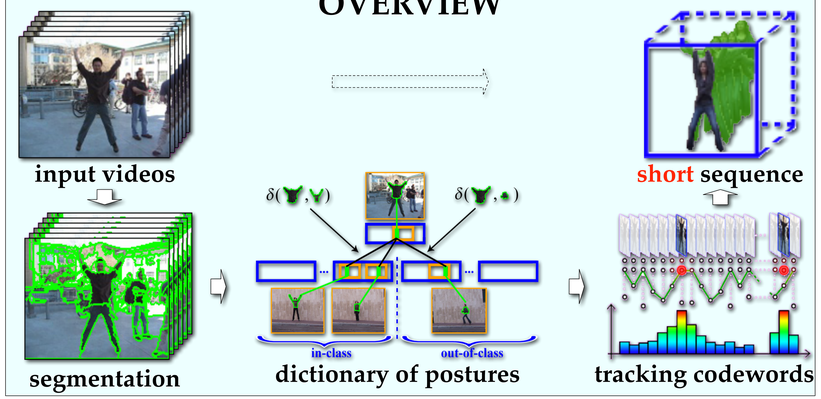
a short sequence of codewords

for exemplar-based recognition.

CONTRIBUTIONS

- Accounting for co-occurrence of actors and tools
- Finding characteristic postures of <u>both</u> actors and tools under weak supervision
- Robust and scalable detection and localization
- Four max-margin methods for dictionary learning
- Proofs of convergence to the global optimum

OVERVIEW



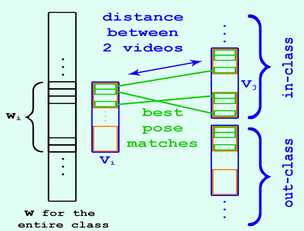
MAX-MARGIN DICTIONARY LEARNING

δ_{ih} , δ_{im} - vectors of distances of all segment of video i to the nearest hit and miss

$\boldsymbol{w}_i$ - vector of weights of all segments of video i

$$\underbrace{\boldsymbol{z} = \begin{bmatrix} \delta_{1m} - \delta_{1h} \\ \vdots \\ \delta_{Nm} - \delta_{Nh} \end{bmatrix}, \quad \boldsymbol{w} = \begin{bmatrix} \boldsymbol{w}_1 \\ \vdots \\ \boldsymbol{w}_N \end{bmatrix}}$$

concatenated vectors of distances and weights of all videos to their nearest misses and hits



maximizing the hypothesis margin of 1-NN classifier

$$\max_{\boldsymbol{w}} \ \boldsymbol{z}^{\mathrm{T}}\boldsymbol{w}, \quad \text{s.t.} \ \ \boldsymbol{w} \geq \mathbf{0}, \ \ \text{and} \ \ \|\boldsymbol{w}\| \leq \gamma \,,$$

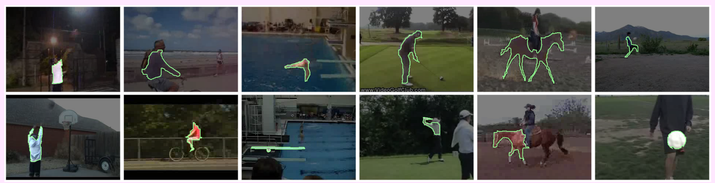
$\Downarrow$

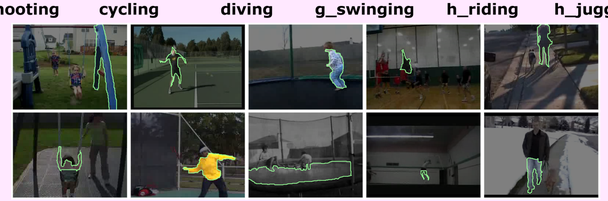
2 alternative formulations, and 2 types of norms: ℓ_1 and ℓ_2

LR:
$$\operatorname*{argmin}_{\boldsymbol{w}} \ \log[1+\exp(-\boldsymbol{z}^{\mathrm{T}}\boldsymbol{w})] + \lambda\|\boldsymbol{w}\|\,, \quad \text{s.t.} \ \ \boldsymbol{w} \geq 0$$

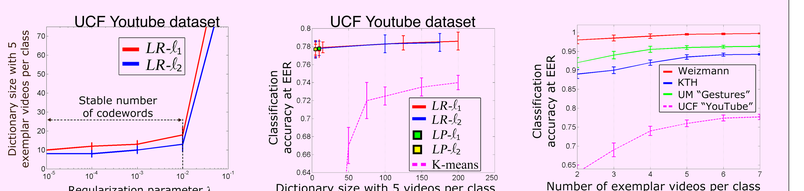
LP:
$$\operatorname*{argmax}_{\boldsymbol{w}} \ \boldsymbol{z}^{\mathrm{T}}\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \quad \text{s.t.} \ \ \boldsymbol{w} \geq 0$$

RESULTS





examples of learned codewords on UCF Youtube [4]



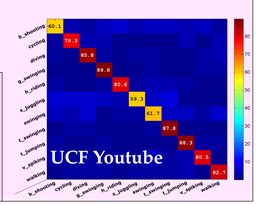
comparison of our 2 formulations and 2 types of norms

	[14]	[12]	[4]	[3]	Ours (LP-ℓ_1)
Weizmann	97.5	X	X	X	99.7
KTH	X	95.7	91.8	87.8	94.2
UM "Gestures"	X	95.2	X	X	96.3
UCF "YouTube"	X	X	71.2	X	77.8

Table 3. Average classification accuracy at EER

	[3]	[8]	Ours (LP-ℓ_1)
pick-up	0.58	0.47	0.60
one-hand wave	0.59	0.38	0.64
jumping jack	0.43	0.22	0.45
two-hands wave	0.43	0.64	0.65

Table 2. AUC for CMU "Crowded" videos



[3] Yao et al. "Learning deformable action templates from cluttered videos," ICCV09
[4] Liu et al. "Recognizing realistic actions from videos in the wild," CVPR09
[8] Ke et al. "Event detection in crowded videos," ICCV07
[12] Lin et al. "Recognizing actions by shape-motion prototype trees," ICCV09
[19] Schueldt et al. "Recognizing human actions: A local SVM approach," ICPR04