On Recognizing Actions in Still Images via Multiple Features

Fadime Sener[1], Cagdas Bas[2], and Nazli Ikizler-Cinbis[2]

[1] Computer Engineering Department, Bilkent University, Ankara, Turkey

[2] Computer Engineering Department, Hacettepe University, Ankara, Turkey

Problem Description

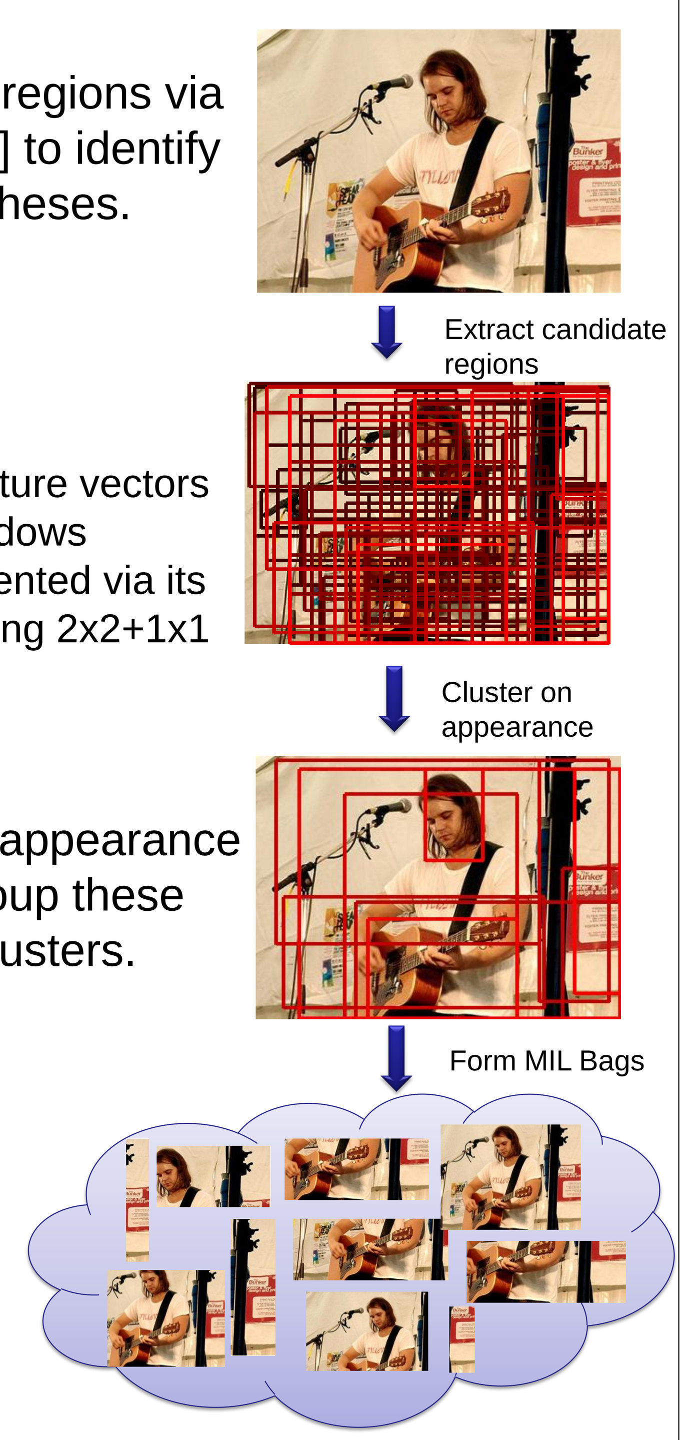
- Still images convey the action information via the pose of the person and the surrounding object/scene context.
- Training explicit object detectors may not be scalable.
- Part/attribute annotation for each action is costly.
- Single features may not be solely reliable and/or discriminative.

Idea

- **Do not use any explicit object detector or part/attribute annotation.**
- **Instead, find candidate regions that contribute to the action recognition and utilize these regions in a weakly supervised manner via Multiple Instance Learning.**
- **Extract many other (possibly noisy) features that are complementary, including**
  - Candidate object region features
  - Facial features
  - Person appearance features
  - Global image features

Multiple Instance Learning for Candidate Object Regions

- Extract visually salient regions via objectness measure [1] to identify candidate object hypotheses.
- Sample 100 windows.
  - Extract dense SIFT feature vectors from each of these windows
  - Each window is represented via its bag-of-words(BoW) using 2x2+1x1 spatial tiling.
- Use k-means over the appearance feature vectors and group these 100 windows into 10 clusters.
- Construct MIL Bags from these candidate regions.



MIL Approach

- We adopt Multiple Instance Learning with Instance Selection (MILES) [2] algorithm for learning the related object regions.
- The similarity between bag B_i and an instance c_l ;

$$s(c_l, \mathrm{B}_i) = \max_j \exp\left(-\frac{D(x_{ij}, c_l)}{\sigma}\right)$$

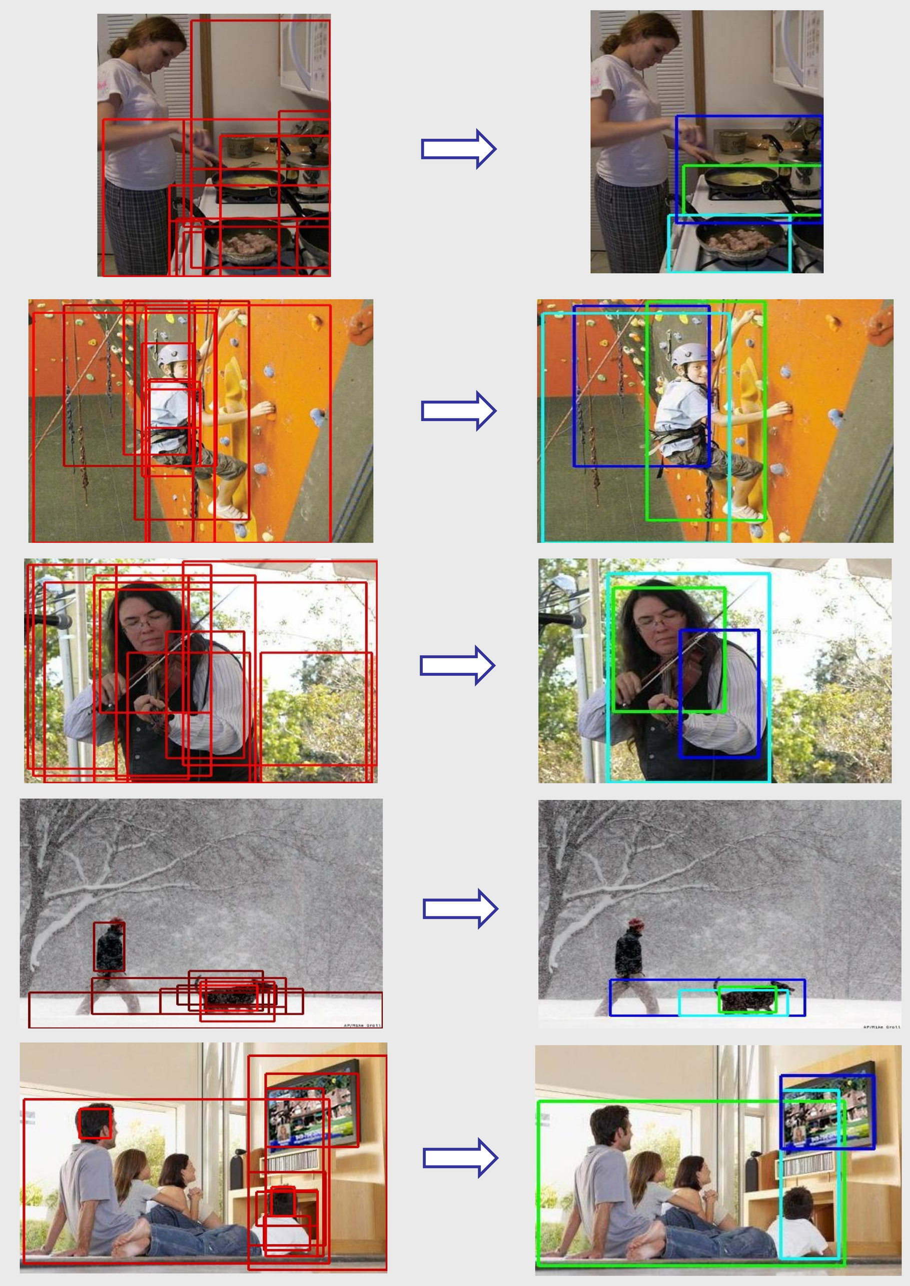
- Each bag represented in terms of its similarities to each of these target concepts (embedding)

$$\mathrm{m}(\mathrm{B}_i) = [s(c_1, \mathrm{B}_i), s(c_2, \mathrm{B}_i), \ldots, s(c_N, \mathrm{B}_i)]^T$$

- Train an L2-regularized SVM with RBF kernel for each action class in a one-vs-all manner.

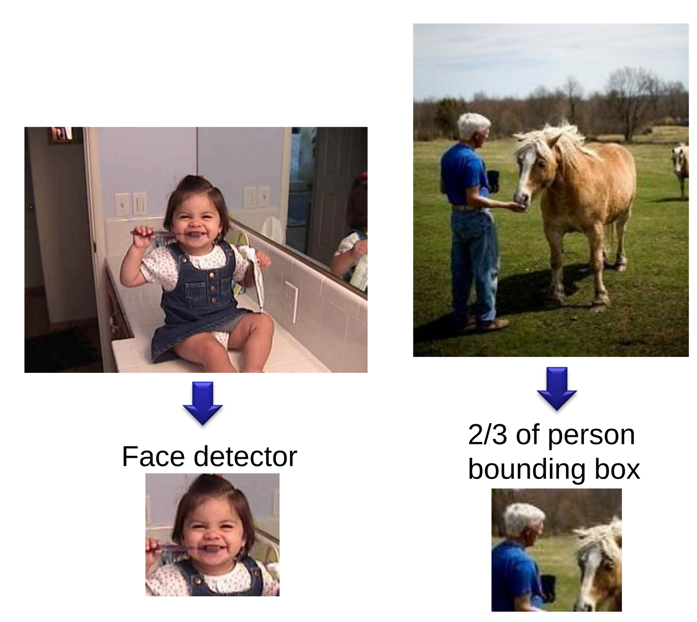
The most contributing concept instances

- 10 candidate object regions (extracted by [1])
- Top 3 regions that contribute to the classification are shown in **green**, **cyan**, **blue**.

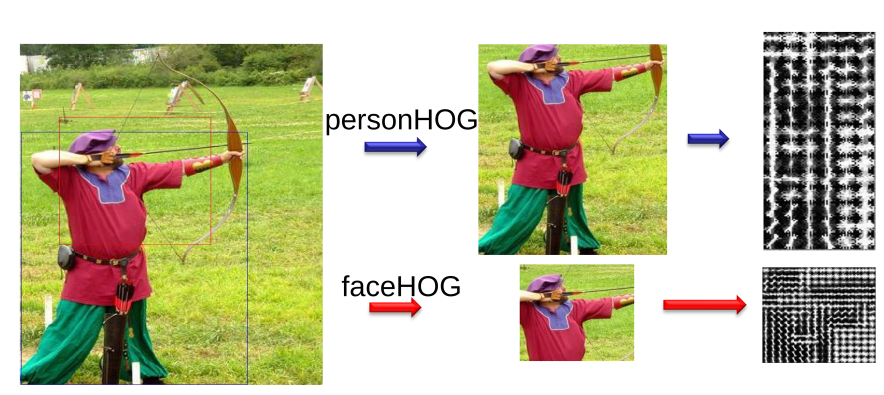


Additional Features

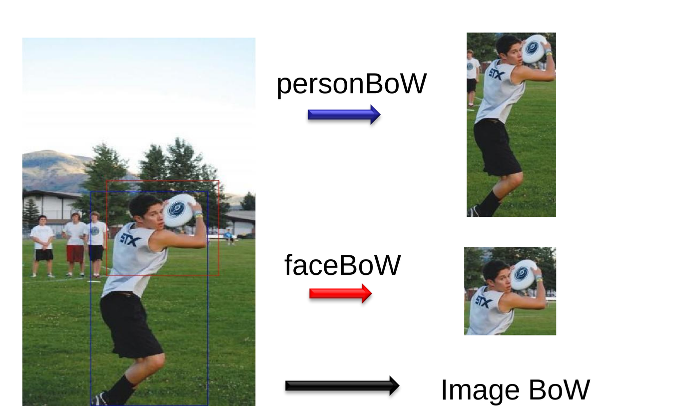
- Facial features and objects around the face can be an indicator of the ongoing action.
- Face regions are extracted with Viola-Jones face detector and dense SIFT features are extracted.
- When face is not detected, top region of the person bounding box is used as the face area.



- Histogram of Oriented Gradient(HOG) features from the person regions in the image



- Bag-of-words(BoW) representations extracted from person, face regions and the whole image.



Effect of the clustering instances in MIL

- We evaluated the effect of the clustering individual instances versus using all instances in the objectness-based MIL formulation.

	accuracy	mAP
objectMIL ($k = 300$)	37.08	34.03
objectMIL ($k = 1000$)	46.78	46.01
objectMIL (no clustering)	**51.34**	**51.80**

- Clustering yields a scalable representation that requires much less time.
- Using all the candidate object regions for instance embedding produces far more effective results in terms of the classification performance.

Experimental Results

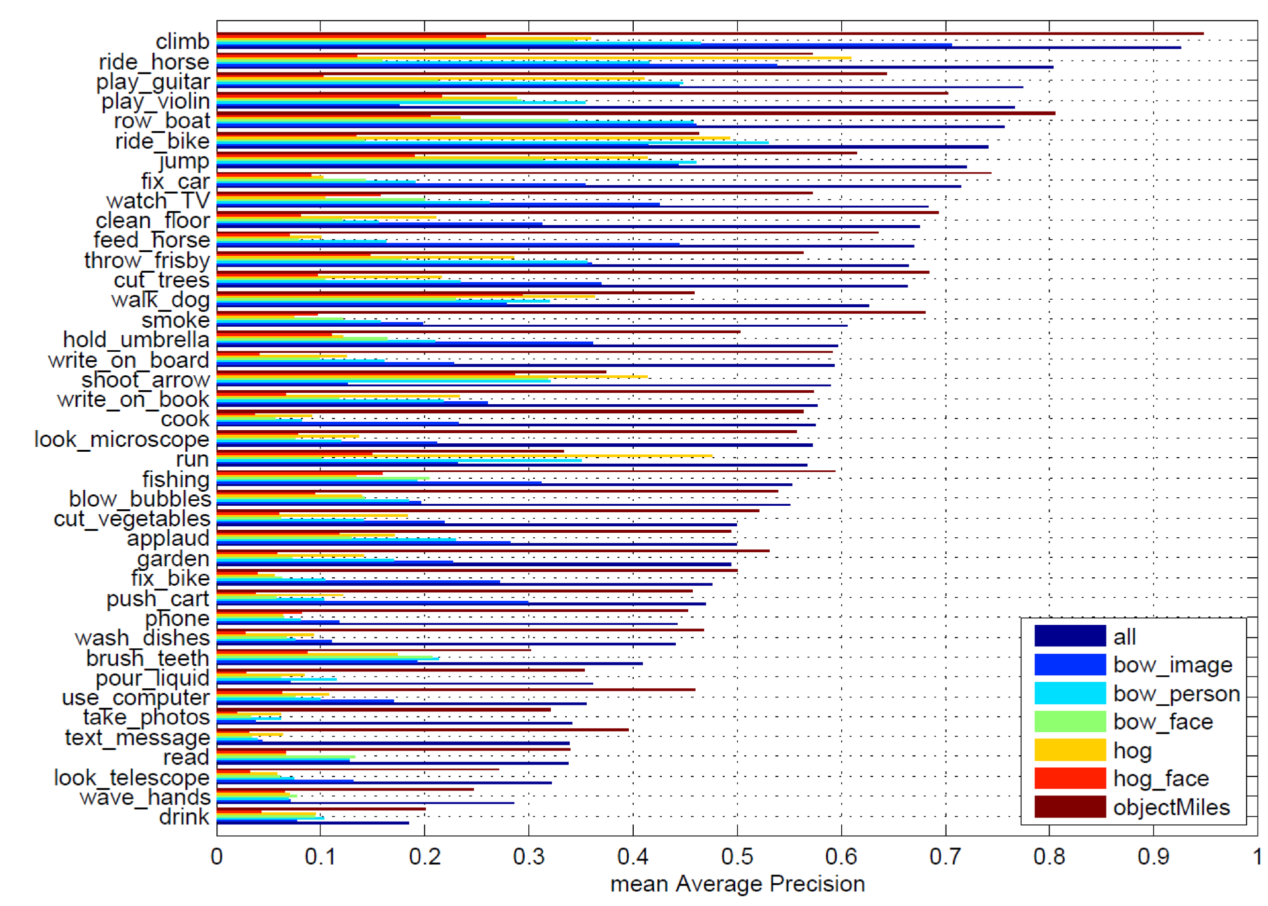
- We evaluate the performance on Stanford 40 Actions, which contains 4000 train images and 5532 test images.

Without explicit object detectors, useful information from the candidate object regions can exracted.

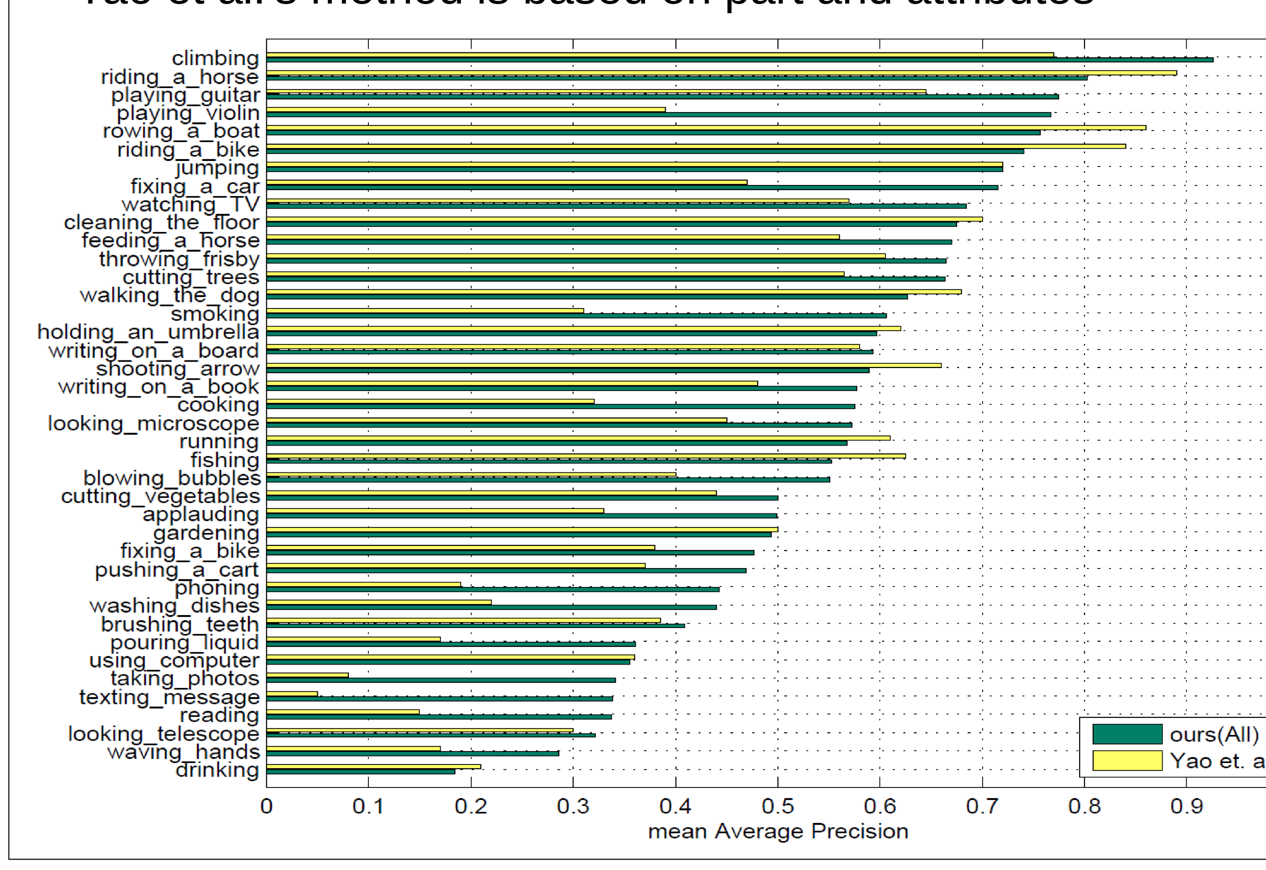
	accuracy	mAP
personHOG	24.75	19.35
personBoW	28.56	21.53
faceHOG	14.01	10.37
faceBoW	17.93	13.83
imgBoW	33.51	26.32
objectMIL	51.34	51.80
imgBoW+objectMIL	52.30	52.23 (No BB)
All(w/o objectMIL)	41.47	36.63
All	**55.93**	**55.55** (With BB)
Yao [5]	NA	45.7

The context information from the whole image is useful for recognizing the action.

- The performance of the individual features wrt each action.
  - Most of the times, the combination works the best.



- Comparison to state-of-the-art method of Yao et al [3] Yao et al.'s method is based on part and attributes



Conclusions and Discussion

- Our experimental results show that the proposed MIL framework is suitable for extracting the relevant object information, without the need for explicit object detectors.
- We have achieved better classification performance compared to the state-of-the-art on the extensive Stanford 40 actions still image dataset.
- Our findings indicate possible future directions, particularly, using richer representations over salient object regions and improving weakly supervised learning of relevant objects/

References

[1] Alexe, B., Deselaers, T., Ferrari, V.: What is an object? In: IEEE Conf. on Computer Vision and Pattern Recognition, San Francisco, USA (2010)

[2] Y. Chen, J. Bi, and J. Z. Wang. Miles: Multiple-instance learning via embedded instance selection. *IEEE TPAMI*, 28(12):1931–1947, 2006.

[3] Yao, B., Jiang, X., Khosla, A., Lin, A.L., Guibas, L.J., Fei-Fei, L.: Human action recognition by learning bases of action attributes and parts. In ICCV, Barcelona, 2011.