**Antonio Hernández-Vela[1,3], Stan Sclaroff[2], Sergio Escalera[1,3]**

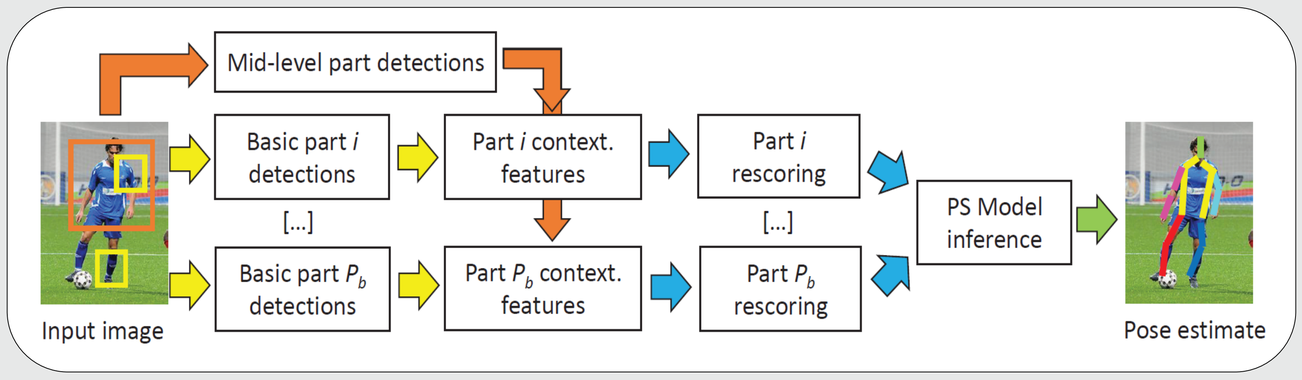
*Universitat de Barcelona[1], Boston University[2] and Computer Vision Center[3]*

# Contextual Rescoring for Human Pose Estimation

## Abstract

A contextual rescoring method is proposed for improving the detection of body joints of a pictorial structure model for human pose estimation. A set of mid-level parts is incorporated in the model, and their detections are used to extract spatial and score-related features relative to other body joint hypotheses. A technique is proposed for the automatic discovery of a compact subset of poselets that covers a set of validation images while maximizing precision. A rescoring mechanism is defined as a set-based boosting classifier that computes a new score for body joint detections, given its relationship to detections of other body joints and mid-level parts in the image. This new score complements the unary potential of a discriminatively trained pictorial structure model. Experiments on two benchmarks show performance improvements when considering the proposed mid-level image representation and rescoring approach in comparison with other pictorial structure-based approaches.

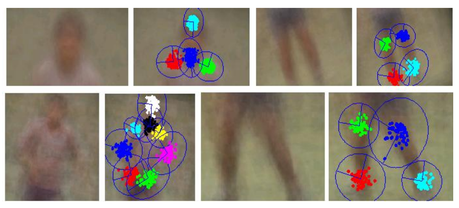
## Human Pose Estimation: Pipeline




## Mid-level part representation

### 1. Poselet [2] training

- Generate seed random windows.
- Procrustes alignment to gather training samples.
- Estimate Gaussian distribution of keypoints.



- Compute keypoint estimation precision

### 2. Poselet selection: weighted set cover

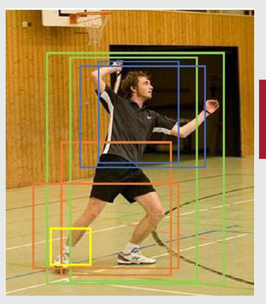
$$\text{minimize} \sum_{\hat{j}} (1 - \text{Prec}(\hat{j}))\mathbf{x}_{\hat{j}}$$

$$\text{subject to} \sum_{\hat{j}:A_{n\hat{j}}=1} \mathbf{x}_{\hat{j}} \geq 1 \; \forall n, \; \mathbf{x}_{\hat{j}} \in \{0,1\},$$

poselet $\hat{j}$
$n$-th validation image

## Contextual Rescoring & Pictorial structure formulation

### 1. Mid-level contextual detections



### 2. Contextual features

| Feature | Value |
|---|---|
| detection score | $[0,\ldots,0,s_{\hat{j}},0,\ldots,0]$ |
| relative position | $(p_i^x - p_{\hat{j}}^x)/he_i, (p_i^y - p_{\hat{j}}^y)/he_i$ |
| relative size | $he_i/he_{\hat{j}}, wi_i/wi_{\hat{j}}$ |
| relative scale | $z_i/z_{\hat{j}}$ |
| distance | $\|(p_i - p_{\hat{j}})\|$ |
| overlap | $(B_i \cap B_{\hat{j}})/(B_i \cup B_{\hat{j}})$ |
| score ratio | $s_i/s_{\hat{j}}$ |
| score difference | $s_i - s_{\hat{j}}$ |

### 3. SetBoost [3] rescoring function

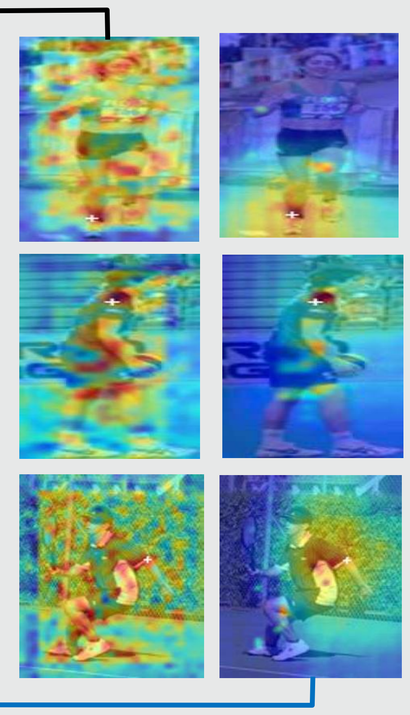
$$R(C) = \sum_{\theta=1}^{\Theta} Q_\theta(C)$$

$$Q_\theta(C) = \alpha_\theta \sum_{c \in C} k_c \cdot q_\theta(c)$$

### 4. Pictorial structure formulation

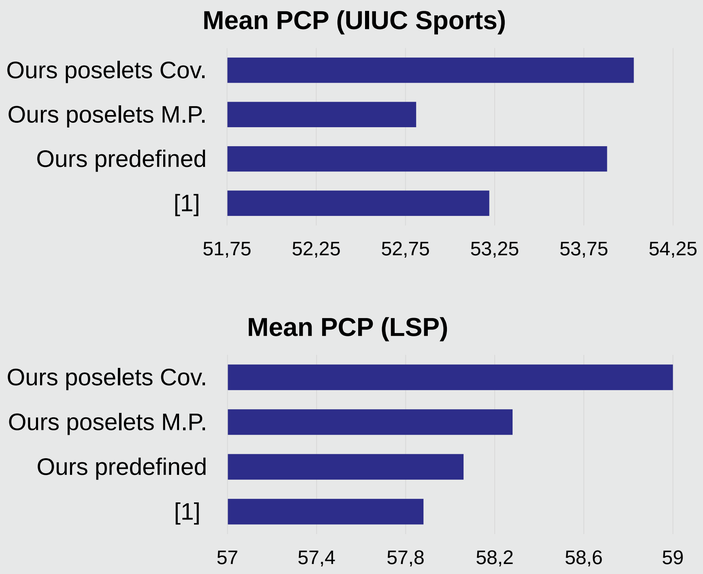
$$S(I,M,p,t) = S(t) + \sum_{i \in V} \left( w_i^{t_i} \cdot \boxed{\Phi(I,p_i)} + \hat{w}_i^{t_i} \cdot \boxed{R_i^{t_i}(C_{B_i}^M)} \right) + \sum_{ij \in E} w_{ij}^{t_i,t_j} \cdot \psi(p_i - p_j)$$

$$S(t) = \sum_{i \in V} b_i^{t_i} + \sum_{ij \in E} b_{ij}^{t_i,t_j}$$



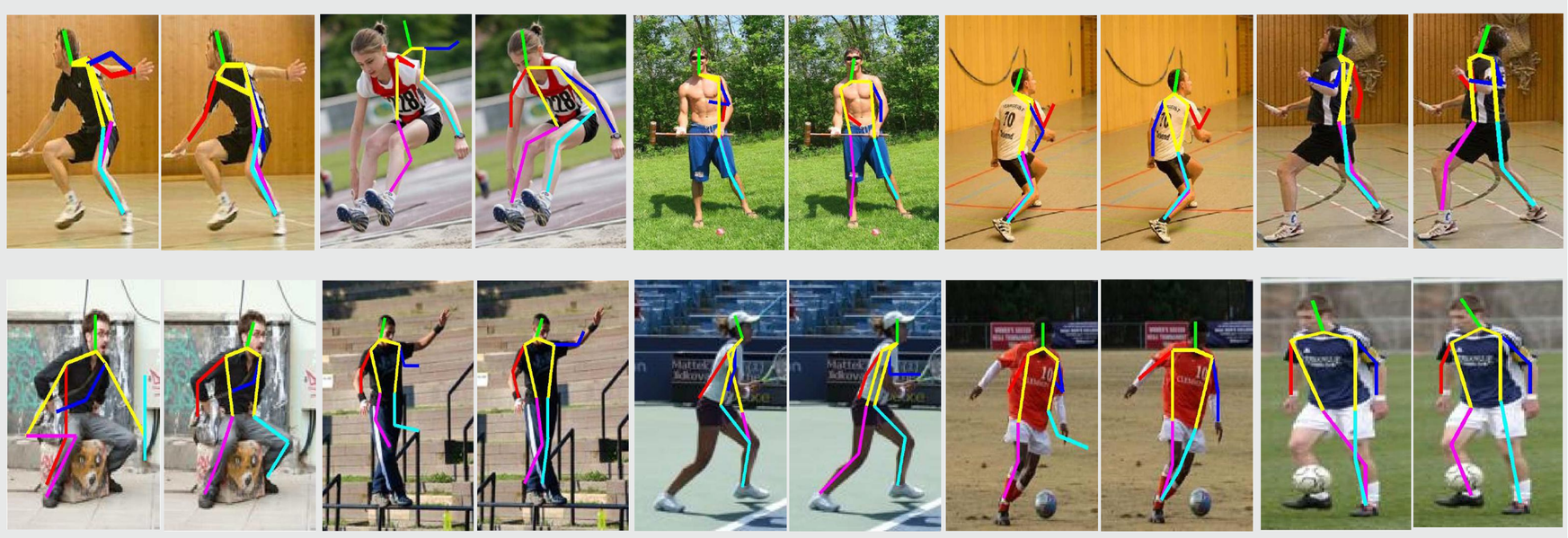
## Results: LSP [4] and UIUC Sports [5] datasets

### Quantitative results




### Qualitative results

(left: Yang & Ramanan [1], right: Ours, poselets cov.)



## References

[1] Y. Yang and D. Ramanan. “Articulated human detection with flexible mixtures of parts”. IEEE TPAMI, 35(12):2878 2890, Dec 2013.
[2] L. Bourdev, S. Maji, T. Brox, and J. Malik. “Detecting people using mutually consistent poselet activations”. In ECCV, volume 6316, pages 168–181, 2010.
[3] R. Gokberk Cinbis and S. Sclaroff. “Contextual object detection using set-based classification”. In ECCV 2012.
[4] S. Johnson and M. Everingham. “Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation”. In BMVC2010.
[5] Y. Wang, D. Tran and Z. Liao. “Learning Hierarchical Poselets for Human Parsing”. In CVPR 2011.