Contextual rescoring for Human Pose Estimation

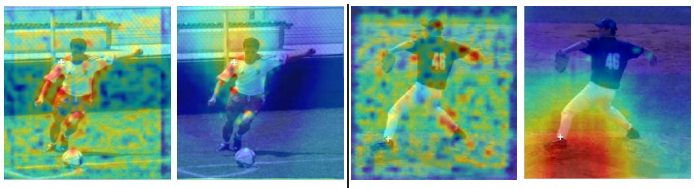
Antonio Hernández-Vela[1]
ahernandez@cvc.uab.cat
Stan Sclaroff[2]
sclaroff@bu.edu
Sergio Escalera[1]
sergio@maia.ub.es

[1] Dept. of Applied Mathematics,
Universitat de Barcelona, Spain
Computer Vision Center, UAB, Spain

[2] Dept. of Computer Science,
Boston University, USA

Given an image of a person, the problem of human pose estimation can be briefly described as localizing the position and orientation of the body limbs. The complexity of the problem comes from issues like background clutter, changes in viewpoint, changes in appearance, self-occlusions of body parts, etc.

Pictorial structures framework has been widely applied in human pose estimationn during the past few years [1]. Yang and Ramanan [7] proposed a simple yet efficient model that outperformed previous state of the art approaches. However, in addition to the difficulties of modelling small image patches for the body joints (see Fig. 1), the performance of their method is also compromised by the use of a tree-structured model. Although trees permit efficient and exact inference on graphical models, the restricted edge structure is insufficient for capturing all the important relations between parts. As a consequence, tree-structured pictorial structures suffer from the so-called "double-counting" phenomena.



(a) (b) (c) (d)

Figure 1: (a) Detection score map for the right shoulder using a classical sliding-window detection approach with a linear SVM trained on HOG features. (b) Rescored version of (a) produced by our context-based rescoring. The original map (a) has a strong score on the actual shoulder location, but also in other regions. Our proposed rescoring produces more spatially-consistent score maps, showing a high response near the correct location, and suppressing false positive locations. In addition, our rescoring method can hallucinate the location of a part, e.g. foot (d) even if there is not a high-scoring region in the original map (c).

In this work, we propose a new method for obtaining robust part detections in a pictorial structure formulation for human pose estimation. Motivated by the fact that small local HOG templates modelling the body joints ("basic parts" from now on) are sensitive to noise, we introduce information from a mid-level representation of the image in order to obtain more reliable basic part detections (see Fig. 2). More specifically, we make the following contributions:

- We introduce a method for the automatic discovery of a compact set of discriminative poselets [2] that offers both high detection precision and a covering of the different poses in a given validation dataset.
- Using this set of poselets as our mid-level image representation, we assign a new score to the detections of a certain basic part through a rescoring function that learns patterns of their contextual relationships.
- We extend the formulation from [7] in order to include the rescored detections.

Experimental evaluation is conducted on two benchmarks: UIUC Sports [6] and Leeds Sports [3]. In the experiments, pose estimation accuracy improves when our proposed rescoring functions are included in the unary potential of a pictorial structure model, using our mid-level part representation (see Fig. 3. In particular, among the different mid-level part representations in our comparative analysis, the automatic discovery

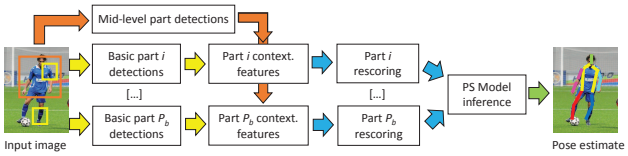


Figure 2: Proposed pipeline for human pose estimation. Given an input image, a set of basic and mid-level part detections is obtained. For each basic part *i* detection, a contextual representation is built based on mid-level part detections, which is used for rescoring the former. The original and rescored detections for all basic parts are then used in inference on a pictorial structure (PS) model to obtain the final pose estimate.

of poselets with covering attains the best results in both datasets. In addition, we report a gain in the pose estimation performance comparable to the one in [4, 5], while reducing the size of the mid-level representation by an order of magnitude (40-50 poselets in our approach vs. more than 1000 in [4, 5]).

Figure 3: Qualitative results for the UIUC Sports dataset (row 1) and LSP dataset (row 2). Leftmost images show the results from [7] and rightmost images show our results.

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, pages 1014–1021. IEEE, 2009.

[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] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, pages 12.1–11, 2010.

[4] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Strong appearance and expressive spatial models for human pose estimation. In *ICCV*, pages 3487–3494, Dec 2013.

[5] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *CVPR*, pages 588–595, 2013.

[6] Y. Wang, D. Tran, and Z. Liao. Learning hierarchical poselets for human parsing. In *CVPR*, pages 1705–1712, 2011.

[7] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(12):2878–2890, Dec 2013.