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

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.

1 Introduction

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.

Among the various methods proposed for human pose estimation, models based on pictorial structures tend to provide superior performance, e.g., recently [1, 13]. The first works on pictorial structures for human pose estimation [6] employed a tree-structured model composed by parts representing the human body (e.g., left foot, lower left leg, etc.), connected following the kinematic constraints of the human body (e.g., left foot is connected to left leg). More specifically, the body parts are modeled as rectangles, parametrized by position, orientation, and size.

In contrast, [18] proposed a discriminatively trained pictorial structure that models the body joints instead of limbs, thus simplifying the formulation and reducing the complexity of inference. Specifically, the body joints are modeled as a mixture of small HOG filters capturing a small neighborhood around them. While attaining better results than previous

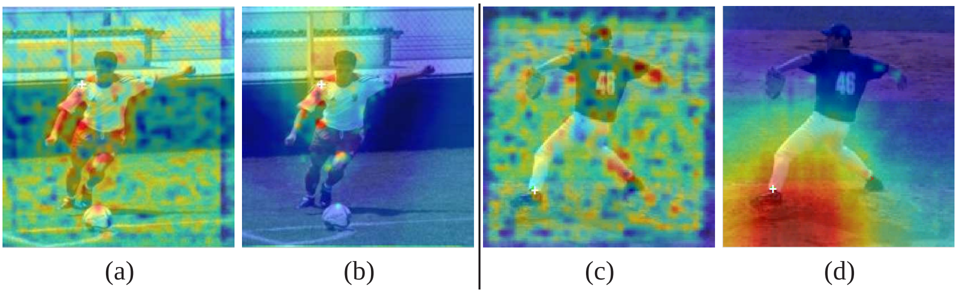


Figure 1: (a) Detection score map for the right shoulder using a classical sliding-window detection approach with a linear SVM trained on HOG features. (b) Rescored version of (a) produced by our context-based rescoring. The original map (a) has a strong score on the actual shoulder location, but also in other regions like the hips. In contrast, with our proposed rescoring, we get a more spatially-consistent score map, showing a high response near the correct shoulder location, and suppression of 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).

works, such small HOG templates can be sensitive to noise and result in several false positive detections at test time, either confusing a part with the background or with another part (see Fig. 1). Moreover, sliding-window detection approaches like this one can fail to recover joints in the presence of occlusions, unless the appearance of an occluded part is explicitly modeled, e.g., by adding a new mixture component.

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 a method for the automatic discovery of a compact set of discriminative poselets [4] that offers both high detection precision and a covering of the different poses in a given validation dataset. Using the evidence of these new mid-level parts, we rescore the basic part detections in order to obtain a more robust basic part detection, and thus improve the inference of human pose.

Experimental evaluation is conducted on two benchmarks: UIUC Sports [17] and Leeds Sports [10]. 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. In particular, among the different mid-level part representations in our comparative analysis, the automatic discovery 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 [11, 12], 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 [11, 12]).

2 Related work

In the context of human pose estimation, Yang and Ramanan [18] 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, 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” phenomenon.

In order to overcome these problems, many extensions of the pictorial structure model have been proposed. Tian and Sclaroff [15] augmented the tree model by adding a small set of edges, and presented an efficient inference algorithm for their model. Dantone, et al. [5] focused on obtaining less noisy appearance-based unary potentials from each part detector. The formulation incorporates color and skin detection, in addition to HOG features. Regressors are employed to estimate the position of a certain joint, given the appearance and location of the other joints. A bridge between human pose estimation and object detection was proposed by Yao and Fei-Fei [2]. They model mutual contextual information between poses and objects for a certain set of human-object interaction activities, like “tennis serve”. The results indiciate that pose estimation can help object detection and vice versa.

Other works introduced higher-level parts in the model, e.g. Poselets [3, 4], to improve the results. In [17], the authors proposed a loopy graph model that incorporates a hierarchical Poselet decomposition of the human body to the existing body parts. In contrast, Pishchulin et al. [11, 12] defined a tree-structured model in which the unary and pairwise terms are conditioned on Poselets evidence. Similarly, Fang and Yi [16] also included mid-level body parts in their tree model, but they propose an algorithm for discovering the best possible tree topology that connects all the parts.

Cinbis and Sclaroff [8] proposed an approach for rescoring detections of different objects, introducing the notion of sets of contextual relations between object detections in an image. Each detection from a certain object class is represented by its context, defined as a set containing detections from every other object detector. After that, a feature vector is extracted from each contextual detection, encoding spatial relations, relative scores and class-related relations. Finally, a generalization of the well-known Adaboost algorithm, called SetBoost, is used for rescoring an object detection given its set-based context representation.

In our work, we recast the pictorial structure formulation from [18] in order to include information from a mid-level representation of the image. More specifically, we follow [12] and define the unary potential of the pictorial structure as a weighted combination of two unary potentials, encoding information from basic and mid-level parts, respectively. However, in contrast to [12], we define our mid-level unary potential as a rescoring function [8], instead of modelling it as a Gaussian distribution. As our mid-level representation, we formulate and test a method for automatic discovery of a compact set of poselets, which maximize precision while enforcing coverage of the poses in a set of validation images.

3 Approach

An overview of the proposed formulation is shown in Fig. 2. We are motivated by the aforementioned limitations of basic, low-level part detectors that are commonly used in pictorial structure models, e.g., HOG patches centered at body joints [18]. In our formulation, we define and learn an additional set of mid-level body part detectors that improve the localization of the basic ones. Mid-level and basic part detectors are computed in order to extract a set of pairwise contextual features between each pair of basic and mid-level part hypotheses. A classifier for a certain basic part class will compute a new score for its detections, based on the set of contextual features computed between the basic and mid-level parts. The original

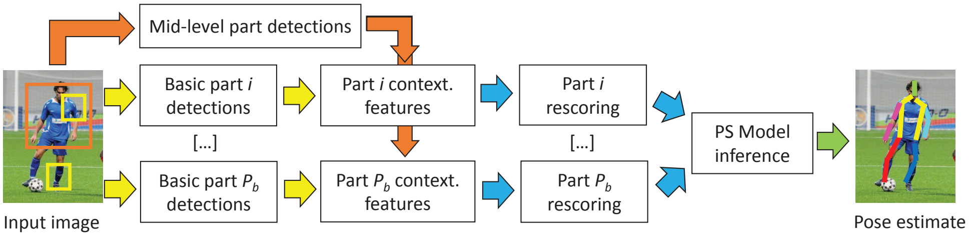

Mid-level part detections
Basic part i detections
[...]
Basic part P_b detections
Part i context. features
Part P_b context. features
Part i rescoring
[...]
Part P_b rescoring
PS Model inference
Input image
Pose estimate

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.

and rescored detections for all basic parts are then used in inference on a pictorial structure model to obtain the final pose estimate.

3.1 Pictorial structure formulation

Let us define $G(V,E)$ as a graph, where V is the set of nodes representing the basic parts in the pictorial structure model, and E is the set of edges connecting them. We define I as an image, p_i as the position of basic part i, t_i as its mixture component, and $B_i = \left[p_i^x \; p_i^y \; wi_i \; he_i \; z_i \; s_i\right]$ as its bounding box including the width wi_i, height he_i, scale z_i and detection score s_i. In addition, $i \in \{1,\ldots,P_b\}$, and $t_i \in \{1,\ldots,K_b\}$. We also define a set of mid-level parts $\hat{j} \in \{1,\ldots,P_m\}$. The score of a pose is given by:

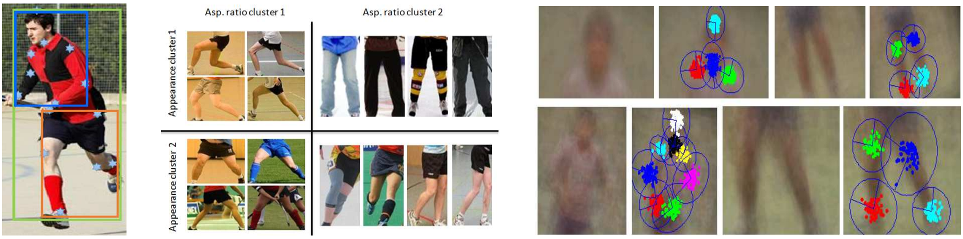
$$S(I,M,p,t) = S(t) + \sum_{i\in V}\left(w_i^{t_i}\cdot\Phi(I,p_i) + \hat{w}_i^{t_i}\cdot R_i^{t_i}(C_{B_i}^M)\right) + \sum_{ij\in E} w_{ij}^{t_i,t_j}\cdot\psi(p_i-p_j), \tag{1}$$

$$S(t) = \sum_{i\in V} b_i^{t_i} + \sum_{ij\in E} b_{ij}^{t_i,t_j}, \tag{2}$$

where $\psi(p_i - p_j) = [dx\; dx^2\; dy\; dy^2]$ encodes the spatial offsets between parts, and $S(t)$ introduces bias factors $b_i^{t_i}$ and $b_{ij}^{t_i,t_j}$ encoding a prior knowledge favoring particular type assignments for part i and particular co-ocurrences of part types, respectively. The unary potential (first summation in Eq. 1) is where we extend the original formulation. In addition to HOG features $\Phi(I,p_i)$ weighted by $w_i^{t_i}$, we define an extra term weighted by $\hat{w}_i^{t_i}$. This new term is defined as a rescoring function $R_i^{t_i} : \mathcal{C} \rightarrow \mathbb{R}$, that receives as input a set of contextual feature vectors $C_{B_i}^M$ associated to a basic part detection B_i and a set $M = \left\{B_{\hat{j}}^n\right\}_{n=1,\hat{j}=1}^{N_m,P_m}$, of mid-level contextual detections, where N_m denotes the number of detections taken from a certain mid-level part $\hat{j}$. In order to simplify the optimization of the model, the basic part detections B_i used for the computation of $C_{B_i}^M$ are obtained by independently trained basic part detectors, using a first initialization of the weights $w_i^{t_i}$ by means of Linear SVM optimization.

3.2 Mid-level part representation

In order to improve basic part detection in the context of a PS model, we define a contextual model based on a set of mid-level body parts. Since higher-level body parts model a larger



(a) (b) (c)

Figure 3: (a) Body joint annotations (blue stars) and different mid-level parts (bounding boxes). (b) Clustering examples for the lower body. (c) Sample Poselet templates. Body joints are shown with colored dots, and estimated Gaussian distributions as blue ellipses.

image portion than just a small local patch as in the case of basic parts, it is expected they will perform better in terms of object detection. We first define a baseline that utilizes a manually-defined set of mid-level parts, which are defined as a hierarchical decomposition of the human body. Then we propose a weighted "set cover" poselet selection method to define a second mid-level representation, which outperforms the baseline in our experiments.

Hierarchical decomposition: We define a mid-level part $\hat{j}$ as a bounding box containing a certain set of body joints, e.g. lower body, upper body, or even the whole body (see Fig. 3(a)). Because of the large variability these parts can have, we also define mid-level parts as a mixture, as in the case of basic parts: we first cluster the training samples $\{I^n_{\hat{j}}\}_{n=1}^N$ into K_{ar} groups based on the aspect ratio $wi^n_{\hat{j}}/he^n_{\hat{j}}$, and apply a second clustering based on appearance following the work from [9]. More specifically, this methodology uses Linear Discriminant Analysis (LDA) to compute what they called "whitened" HOG features a.k.a. WHO features. The main advantage of these features is that they obtain more visually meaningful clustering results than simple HOG features. Finally, Ncuts [14] is applied with a certain K_{app} value indicating the number of clusters we want to create. After performing both clustering steps, the total number of mixture components for a mid-level part becomes $K_m = K_{ar} \cdot K_{app}$.

Poselet-based parts: In this case, we use a similar methodology to the one proposed by Bourdev et al. [4] to define our mid-level parts $\hat{j}$. We generate a large number (thousands) of random seed windows $B^n_{\hat{j}}$ from the training set images $\{I^n\}_{n=1}^N$, and for each one of them we collect similar patches from other training images by Procrustes alignment on the body joint annotations p^{gt}_k from the ground-truth. For each seed window and its associated set of similar examples, we train a mid-level part detector $\omega_{\hat{j}}$. Additionally, we model the spatial distribution of the keypoints k that fall inside each seed window as Gaussian distributions $(\mu^k_{\hat{j}}, \Sigma^k_{\hat{j}})$, that we use to look for True Positives (TP) and False Positives (FP) when testing each detector $\omega_{\hat{j}}$ in a validation set. In order to do that, we use the same criterion as the PCP metric, widely used for evaluating human pose estimation methods. More specifically, we consider a detection as a TP if:

$$\mathrm{dist}(\mu^k_{\hat{j}}, p^{gt}_k) \leq \kappa, \forall k \in B_{\hat{j}}, \tag{3}$$

where κ is a threshold value, i.e. we classify a detection as a TP if the distance between the body joint estimations $\mu^k_{\hat{j}}$ and their corresponding ground-truth annotations p^{gt}_k is below a threshold κ, for all the joints k contained in the poselet. On the contrary, we consider a FP

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

Table 1: List of contextual features included in $c_{B_i,B_{\hat{j}}}$.

if none of the keypoints k fulfill the condition above. Since the seed windows are generated randomly, some of them will be redundant, or some others might have poor performance, so we need to select a subset of relevant poselets. This selection is treated as a “set cover” problem in [4]; poselets are selected in a greedy manner so as to “cover” more examples, i.e. the poselets that found TP detections in a larger number of training images. However, this methodology does not prioritize poselets with good performance if they only fire in a little subset of training images. In order to overcome this problem, we propose using a weighted version of the “set cover” problem, in which the precision of the selected poselets is maximized, while ensuring coverage of the images in a validation dataset. We define a binary matrix $A_{n\hat{j}}$ to keep track of which poselet $\hat{j}$ fires in which n-th validation image. Finally, we formulate this weighted “set cover” problem with the following integer programming:

$$\text{minimize} \sum_{\hat{j}} (1 - \text{Prec}(\hat{j}))\mathbf{x}_{\hat{j}} \quad \text{subject to} \sum_{\hat{j}:A_{n\hat{j}}=1} \mathbf{x}_{\hat{j}} \geq 1\ \forall n,\ \mathbf{x}_{\hat{j}} \in \{0,1\}, \tag{4}$$

where $\text{Prec}(.)$ computes the precision of a poselet. The solution $\mathbf{x}$ finds the subset of poselets $\{\hat{j}\}$ s.t. $\mathbf{x}_{\hat{j}} = 1$, i.e. a set of poselets ensuring that in every validation image there is at least one poselet that fires. The constraints of the integer program enforce each validation image n to be covered by at least one poselet, but also the best-performing ones are prioritized, since we are minimizing $(1 - \text{Prec}(\hat{j}))$. In order to find the solution, we use a Linear Programming relaxation $(\mathbf{y}_{\hat{j}} \in \mathbb{R}_{\geq 0}, \mathbf{y}_{\hat{j}} \leq 1)$ and round the solution $\mathbf{y}$ to obtain $\mathbf{x}$.

Classifier training: The same LDA-based framework we used in our baseline mid-level representation also allows us to train a different detector for each mid-level part $\hat{j}$, much faster than using a classical SVM framework. More specifically, learning a classifier for a certain part $\hat{j}$ is as simple as computing the mean HOG vector $\mu_{\hat{j}}$ among the samples in it. Since Σ and μ_0 are related to the negative set of samples, they are just computed once, and reused for learning the classifiers for all clusters. Finally, these detectors are then run over the images, obtaining the set of detections M used for computing $C_{B_i}^M$.

3.3 Contextual rescoring

We build our contextual model on top of the mid-level part representation presented in Section 3.2. More specifically, we want to model underlying spatial and score-related relationships between basic and mid-level part detections. By doing this, a certain mid-level part detection would be able to determine a hypothesis for the location of a certain basic part. For this task, we define the context of a given basic part detection B_i as a set $C_{B_i}^M = \left\{ c_{B_i,B_{\hat{j}}} \mid \forall B_{\hat{j}} \in M \right\}$, composed by contextual feature vectors $c \in C$. These contextual feature vectors c encode spatial, score-related and class-related relationships between a reference basic part detection B_i and a contextual mid-level detection $B_{\hat{j}}$. We use the same set

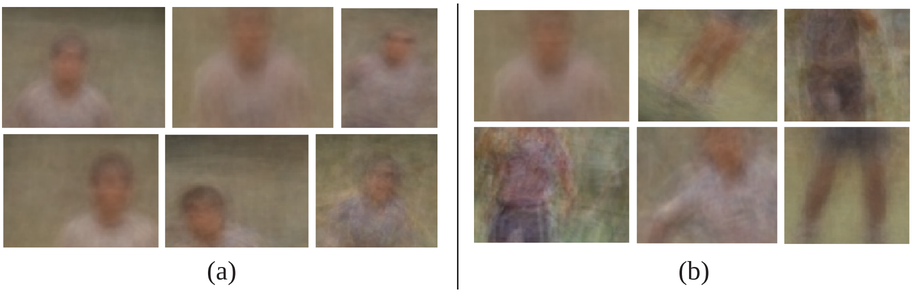


Figure 4: Sample poselets from UIUC Sports dataset. (a) Poselets with highest precision. (b) Poselets discovered by our selection method, maximizing precision and enforcing covering.

of features as [8], with the addition of the relative scale. The specific set of features we use is summarized in Table 1. Finally, the rescoring function given a set of contextual feature vectors C is then defined as:

$$R(C) = \sum_{\theta=1}^{\Theta} Q_\theta(C), \quad Q_\theta(C) = \alpha_\theta \sum_{c \in C} k_c \cdot q_\theta(c), \tag{5}$$

where Q_θ is a weak set classifier, and q_θ is a weak item classifier, weighted by α_θ. The term k_c introduces an additional weight related to the relevance of the item. In practice, k_c is set to its corresponding detection score $s_{\hat{j}}$, and q_θ functions are defined as decision trees with F leaves, which generate $U_1, \ldots, U_F$ partitions of the feature space. The weights α^f for each leaf f are computed following the SetBoost algorithm [8].

In order to train the rescoring function $R_i^{t_i}$ for basic part i and type t_i, we run its corresponding basic part detector on a set of images, as well as the whole set of mid-level part detectors $\hat{j}$. Then, for each basic part detection B_i, we compute the corresponding mid-level contextual feature set $C^M_{B_i}$, and assign a binary label $y_{B_i} \in \{-1, 1\}$, whether the overlapping $O(B_i, B_j) = B_i \cap B_j / B_i \cup B_j$ is below or above a threshold value τ. The complexity of rescoring a basic part detection is $\mathcal{O}(|M|)$.

4 Experiments

In order to present the results, we first describe the data and validation procedure used in our experiments, the different methods and parameters, and evaluation measures.

Data: We conducted experiments with two publicly available challenging datasets: UIUC Sports [17], which contains $1,299$ annotated images of people playing 18 different sports, and Leeds Sports (LSP) [10], which comprises by $2,000$ images of people playing 8 different sports. The annotations for both datasets consist of 14 position labels, one for each body joint: left/right2 ankle, knee, hip, wrist, elbow and shoulder, neck and head top. In the case of LSP, the annotations are observer-centric, i.e. left/right labels on the limbs are defined as the left-most/right-most limb in the image respectively. In contrast, the labels in UIUC Sports are person-centric, i.e. left/right labels are related to the actual left/right limbs of the person in the image. We divided each one of these datasets into three subsets, namely training, validation and test. The training set contains 50% of the images and is used for

(a) UIUC Sports dataset

Method	Torso	Upper Leg		Lower Leg		Upper Arm		Forearm		Head	Mean
YR [18]	**85.62**	62.12	56.25	**60.62**	49.06	40.94	45.00	26.25	29.06	76.56	53.22
Ours-predefined	85.31	62.50	**60.00**	58.75	49.69	41.25	**47.50**	**26.56**	29.06	**78.13**	53.88
Ours-poselets M.P.	85.00	**63.12**	57.81	60.00	**51.25**	40.00	44.69	23.44	27.19	75.62	52.81
Ours-poselets cov.	**85.62**	62.50	59.06	59.06	50.62	**43.75**	**47.50**	26.25	**30.31**	75.62	**54.03**

(b) LSP dataset

Method	Torso	Upper Leg		Lower Leg		Upper Arm		Forearm		Head	Mean
YR [18]	83.00	66.40	67.60	63.60	**62.20**	50.00	49.20	31.00	29.20	76.60	57.88
Ours-predefined	82.80	66.40	**68.00**	65.00	62.00	51.40	47.00	30.00	**30.40**	77.60	58.06
Ours-poselets M.P.	83.60	67.80	67.80	**65.20**	61.20	51.20	49.00	30.00	29.80	77.20	58.28
Ours-poselets cov.	**83.80**	**69.80**	67.40	**65.20**	61.80	**53.60**	**50.60**	**31.20**	28.60	**78.00**	**59.00**

Table 2: Comparison of pose estimation results (PCP) for different mid-level representations.

learning the PS model. The validation set contains 25% of the images and is used in learning the rescoring functions $R_i^{t_i}$. Finally the test set contains the remaining 25% of images.

Method and validation: For our manually-defined mid-level representation, we decompose the human body into three parts: full body, upper body and lower body (see Fig. 3(a) for an example of these parts). More specifically, we set $K_{ar} = 5$ and $K_{app} = 3$, so we have a total of $3 \times K_{ar} \times K_{app} = 45$ mid-level detectors in our manually-defined mid-level representation. In practice, some aspect ratios for certain parts are too skewed so we discard them and get a total of $P_m = 42$ mid-level detectors. We chose these parameters in order to keep our contextual model as simple as possible, but being able to capture the variabilities present in the data. Our poselet selection method automatically selects 43 and 50 poselets in the UIUC Sports and LSP datasets respectively, from an initial set of $2,000$ poselet proposals (see Fig. 4). In order to define the set M of contextual detections, we take the $N_m = 2$ best detections from each mid-level part detector. Each rescoring function $R(C)$ is defined as a forest of $\Theta = 20$ decision tree weak classifiers, each one of them having a maximum number $V = 150$ of leaf nodes. In addition, we use $\lambda = 0.01$ and $\tau = 0.6$.

Evaluation measurement: We use the Percentage of Correctly-placed Parts (PCP) [7] as the evaluation measure, like most recent works in human pose estimation.

Human pose estimation: Our PS model for these experiments is composed of $P_b = 14$ different parts, where each part is defined as a mixture with $K_b = 6$ components. Table 2 shows a performance comparison of the original PS model from Yang and Ramanan [18] against our rescoring-based extension, with three different manually-defined mid-level representations: (1) the hierarchical decomposition defined at the beginning of Sec. 3.2 (Ours-predefined), (2) a set of P_m poselets maximizing precision (Ours-poselets M.P.) and (3) the set of poselets obtained with our selection algorithm (Ours-poselets cov.). In case (2) we fix $P_m = \sum_{\hat{j}} \mathbf{x}_{\hat{j}}$, which is the number of poselets selected by (3). Our poselet selection method that enforces a covering performs the best in both datasets, with a PCP improvement of +0.81% and +1.12% in UIUC Sports and LSP datasets, respectively. In both datasets, the main improvement comes from the upper arms (+2.8% and +2.5% in UIUC Sports, +3.6% and +1.4% in LSP), while other parts perform equal or slightly worse. In the case of the LSP dataset we attain somewhat better performance improvement than in the UIUC Sports dataset; in addition to the great improvements in the upper arms, we also obtain PCP increases of +3.4% for the upper legs, +1.6% for lower legs, and +1.4% for the head (all w.r.t. [18]). This could be explained by the fact that ground-truth annotations for LSP are

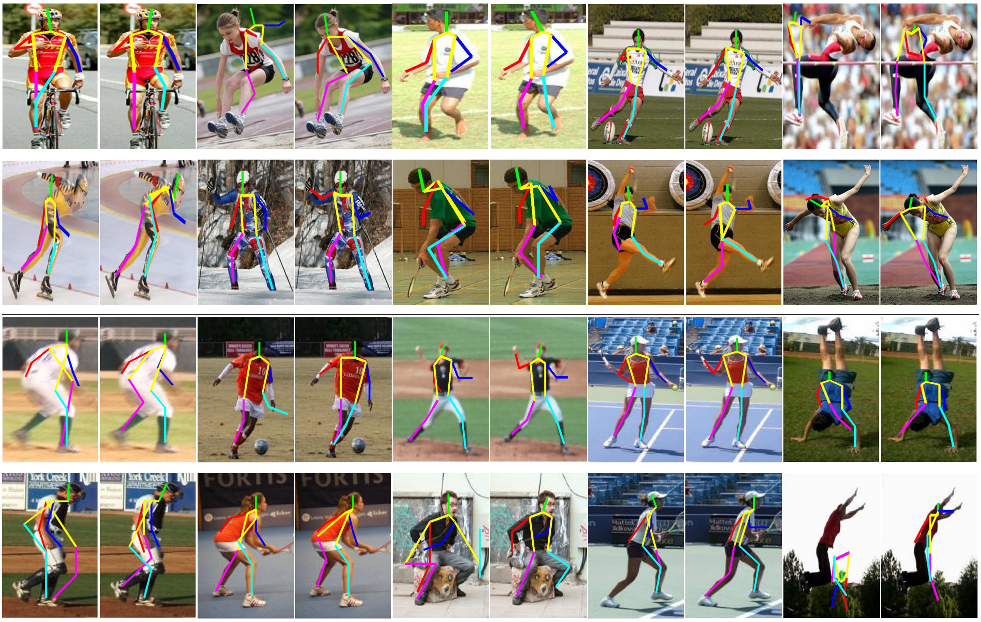


Figure 5: Qualitative results for the UIUC Sports dataset (rows 1-2) and LSP dataset (rows 3-4). Left image from each pair of images shows the result from [18] and right image shows our results. Last column show failure cases.

observer-centric, so our rescoring functions can generalize better the relative locations of left/right basic parts with respect to mid-level parts.

Fig. 5 shows some qualitative results in both datasets [1]. As we can see, the proposed method can better localize the arms in some cases where [18] fails. Additionally, our method can also recover from the "double counting" problem in some other cases. However, there are still some hard cases where our method cannot fully succeed in estimating a pose that matches the ground-truth (see last column in Fig. 5). In order to solve this without increasing the number of mid-level parts, we would need to run our mid-level detectors at different orientations, in order to capture large rotations of the human body.

Finally, we retrained and tested the PS model from Pishchulin et al. [11] with and without their mid-level Poselet representation. When including Poselet information, they obtain a PCP improvement of $+1.06\%$, comparable to our $+1.12\%$. It is worth to note that their mid-level representation is formed by ~ 1000 poselets, while ours contains just 50 poselets.

5 Conclusion

We have proposed a contextual rescoring methodology for improving human pose recovery [2]. In order to obtain more accurate basic part detections, we use a contextual rescoring mechanism based on detections of higher level body parts. We define a simple and compact mid-level body part representation modelling each mid-level part as a mixture, clustering the

[1] Additional qualitative results and rescored maps can be found in the supplementary material.

[2] Our implementation is available at `http://www.cvc.uab.cat/~ahernandez/contextual.html`.

samples using aspect ratio and appearance features. In addition, we propose a method for automatically discovering a set of discriminative poselets for a richer mid-level representation. Using spatial and score-related features extracted from a set of mid-level part detections, we rescore the body joints hypothesis and combine them with the original scores in the unary potential of a PS model.

The experiments with two standard benchmarks demonstrate that by including contextual information from mid-level part detections, we can obtain a better part localization, especially for joints with a more constant relative position among the mid-level parts. Moreover, when poselets are chosen so as to cover a validation set during training using our proposed formulation, experiments show that it is possible to get PCP performance gains comparable to the ones of [11, 12], while using substantially fewer poselets in our model (around 50 in our model vs. more than 1000 in the model of [11, 12]).

Acknowledgements

This work has been partly funded by grants #1029430 and #0910908 from the US National Science Foundation, and research project ref. TIN2013-43478-P from the Spanish Ministry of Economy and Competitiveness . The work of Antonio is supported by an FPU fellowship from the Spanish government.

References

[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] Y. Bangpeng and F. Li. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, pages 17–24, 2010.

[3] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *CVPR*, pages 1365–1372, 2009.

[4] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *ECCV*, volume 6316, pages 168–181, 2010.

[5] M. Dantone, J. Gall, C. Leistner, and L. Van Gool. Human pose estimation using body parts dependent joint regressors. In *CVPR*, pages 3041–3048, 2013.

[6] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *Int. J. Comput. Vision*, 61(1):55–79, January 2005.

[7] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, pages 1–8, 2008.

[8] R. Gokberk Cinbis and S. Sclaroff. Contextual object detection using set-based classification. In *ECCV*, volume 7577, pages 43–57, 2012.

[9] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. In *ECCV*, volume 7575, pages 459–472, 2012.

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

[11] 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.

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

[13] B. Sapp, C. Jordan, and B. Taskar. Adaptive pose priors for pictorial structures. In *CVPR*, pages 422–429, 2010.

[14] J. Shi and J. Malik. Normalized cuts and image segmentation. *PAMI*, 22(8):888–905, 2000.

[15] T. Tai-Peng and S. Sclaroff. Fast globally optimal 2d human detection with loopy graph models. In *CVPR*, pages 81–88, 2010.

[16] F. Wang and Y. Li. Beyond physical connections: Tree models in human pose estimation. In *CVPR*, pages 596–603, 2013.

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

[18] 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.