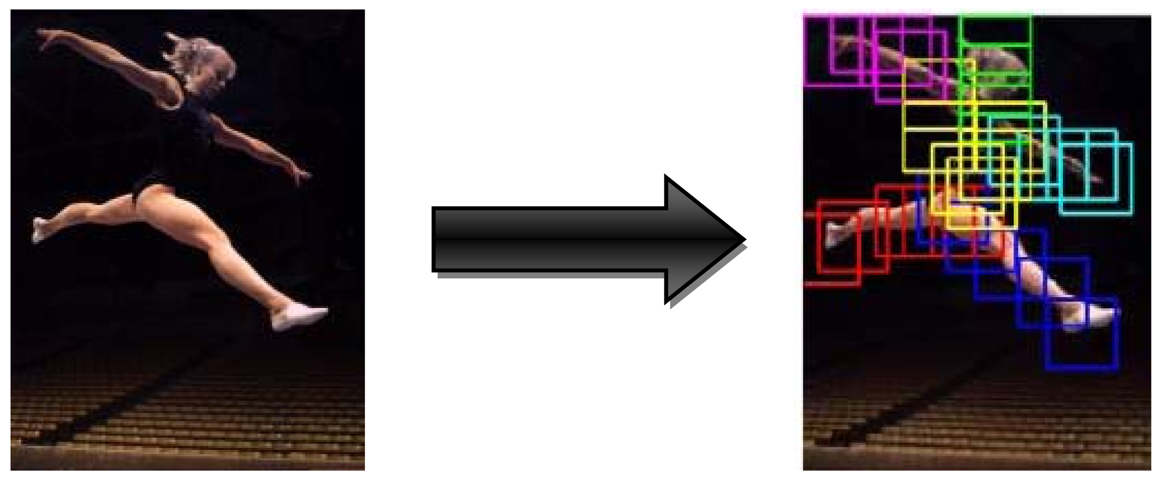
Articulated Pose Estimation with Flexible Mixtures of Parts

Yi Yang, Deva Ramanan

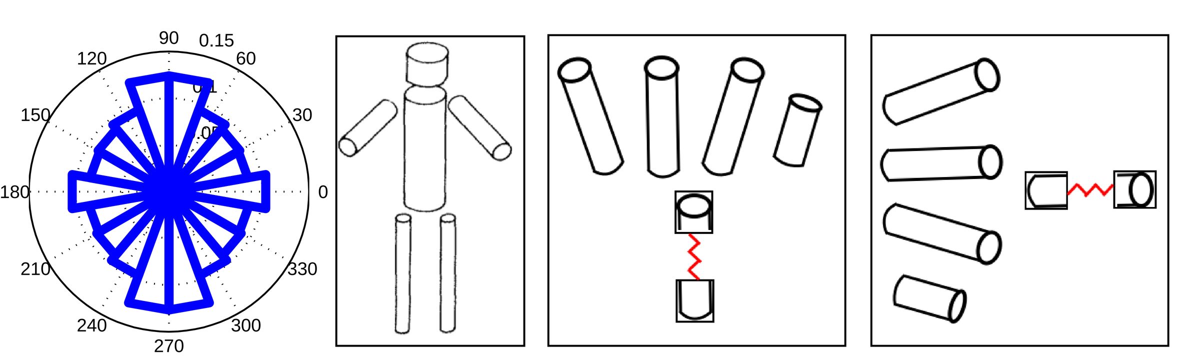
Department of Computer Science, University of California, Irvine

Introduction



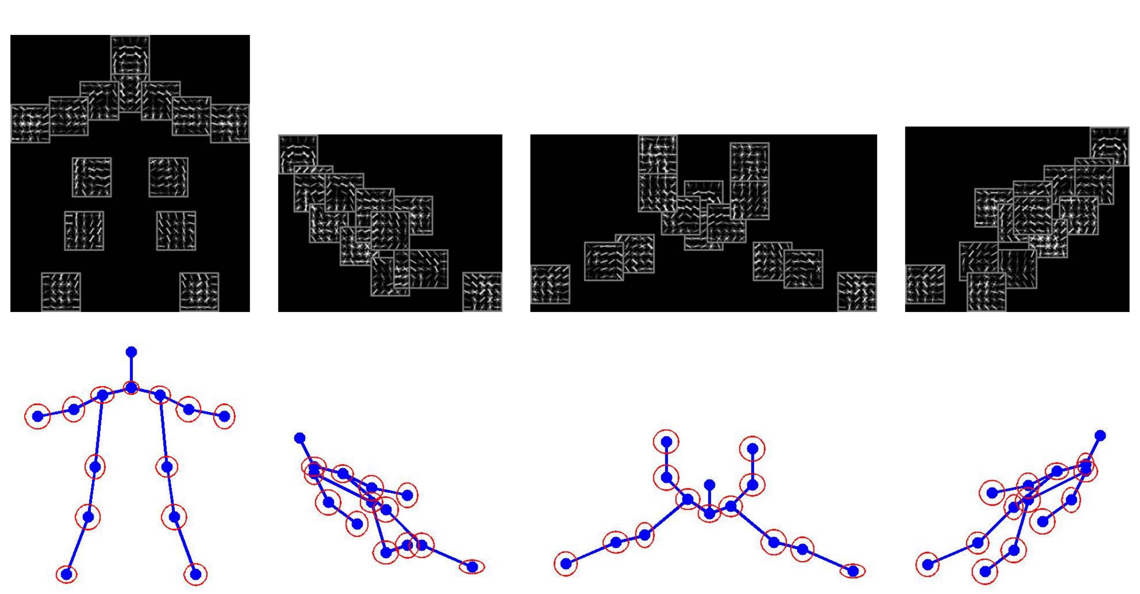
We describe a new method for human pose estimation in static images based on a novel representation of part models, ourperforming past work while being orders of magnitude faster.

Motivation



- Classic articulated limb model **(middle left)** for full-body pose estimation is difficult because limbs vary greatly in appearance due to changes in clothing and body shape as well as changes in viewpoint manifested in in-plane orientations and foreshortening.
- Articulated limb models obtained by rotating single template may be sub-optimal since they cannot exploit orientation-specific background statistics, due to the fact that natural images contain more horizontal edges than vertical and diagonal edges **(left)**.
- We address these problems by introducing a mixture of non-oriented pictorial structures **(middle right, right)** that deform to model a family of affinely-warpped templates.

Model Visualization



A visualization of our full-body model trained on the Parse dataset. We show them as 4 separate models, but we emphasize that our representation allows for the composition of any part type with any other part type, where the score associated with each combination decomposes into a tree (and so is efficient to search over) and is learned from training data.

Model

We augement the standard pictorial structure model:

$$S(x,l,k) = \Sigma_{i\in V}\left[\, w^i_{k_i}\cdot\phi(x,l_i) + b^i_{k_i}\,\right] + \Sigma_{ij\in E}\left[\, w^{ij}_{k_i,k_j}\cdot\psi(l_i,l_j) + b^{ij}_{k_i,k_j}\,\right]$$

- x : image window
- l_i : the pixel location of part i
- k_i : the type (mixture component) of part i, our motivating example of types include orientations of a part but types may span semantic classes
- $\phi(x,l_i)$: local appearance feature (e.g. HOG) extracted from location l_i
- $\psi(l_i,l_j)$: spatial feature extracted from the relative location l_i w.r.t. l_j
- $w^i_{k_i}$: local appearance template for part i with type assignment k_i
- $b^i_{k_i}$: local appearance bias for part i with type assignment k_i
- $w^{ij}_{k_i,k_j}$: spatial spring parameter for pair of types (k_i,k_j)
- $b^{ij}_{k_i,k_j}$: the bias for co-occurrences of pair of types (k_i,k_j)

Inference

Inference corresponds to maximizing $S(x,l,k)$ over l and k. When the relational graph (V,E) is a tree, this can be done efficiently with dynamic programming. Let kids(i) be the set of children of i in (V,E). We compute the message of part i passes to its parent j :

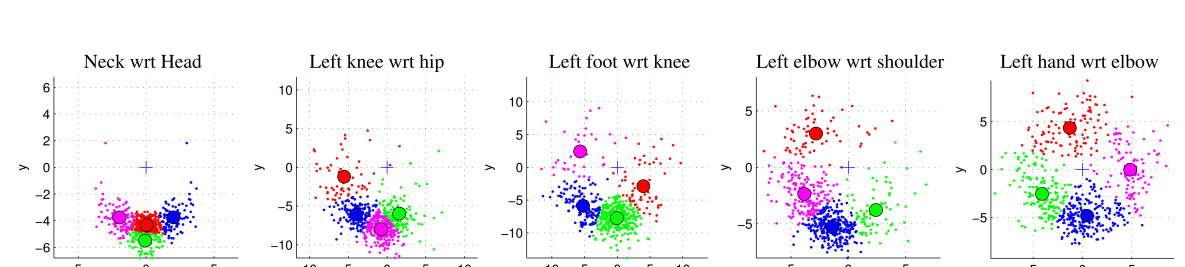
$$s_i(l_i,k_i) = b^i_{k_i} + w^i_{k_i}\cdot\phi(x,l_i) + \Sigma_{j\in \text{kids}(i)} m_j(l_i,k_i)$$
$$m_i(l_j,k_j) = \max_{k_i} b^{ij}_{k_i,k_j} + \max_{l_i} s_i(l_i,k_i) + w^{ij}_{k_i,k_j}\cdot\psi(l_i,l_j)$$

Learning

Given labeled positive examples $\{x_n,l_n,k_n\}$ and negative examples $\{x_n\}$, we write $z_n=(l_n,k_n)$, and $S(x,z)=\beta\cdot\Phi(x,z)$. We learn the model using structural SVM :

$$\begin{aligned} &\arg\min_{\beta,\xi_n\ge 0} && \frac{1}{2}||\beta|| + C\Sigma_n\xi_n \\ &\text{s.t.}\quad \forall n\in\text{pos} && \beta\cdot\Phi(x_n,z_n)\ge 1-\xi_n \\ &\qquad\forall n\in\text{neg},\forall z && \beta\cdot\Phi(x_n,z)\le -1+\xi_n \end{aligned}$$

Partial Supervision



Most human pose datasets include images with labeled joint positions. We define parts to be located at joints so these provide part locations l. We assume part types k correspond to different relative locations of a part with respect to its parent in the relational graph (V,E). We use K-means for type initialization and treat the type as a latent variable that is optimized by coordinate descent during learning.

Results

Image Parse Testset

Method	Torso	Head	U.leg	L.leg	U.arm	L.arm	Total
R Grad[3]	39.5	21.4	20.7	20.7	12.7	11.7	19.2
R Grad+RGB[3]	52.1	37.5	31.0	29.0	17.5	13.6	27.2
ARS HOG[4]	81.4	75.6	63.2	55.1	47.6	31.7	55.2
JE HOG[5]	73.2	62.4	58.6	52.2	47.8	32.5	51.8
JE HOG+RGB[5]	77.6	68.8	61.5	54.9	53.2	39.3	56.4
SNH ROG+RGB[6]	**91.2**	76.6	71.5	64.9	50.0	34.2	60.9
JE NLHOG[7]	85.4	76.1	73.4	65.4	**64.7**	**46.9**	66.2
Our Model HOG	89.8	**87.8**	**78.5**	**69.0**	64.4	36.1	**67.4**

- We compare our model to all published results on the Parse dataset, using the standard criteria of PCP [8]. We beat all previous results on both total and per-part basis, except for torso and lower arm detection.
- [5] uses the same HOG feature set as us but embedded in a classic articulated pictorial structure. The relative improvement of our approach is 20%.

Subset of Buffy Testset

Method	Torso	Head	U.arm	L.arm	Total
TF[9]					62.3
ARS[4]	90.7	95.5	79.3	41.2	73.5
EFZ[10]	98.7	97.9	82.8	59.8	80.1
SJT[11]	100	100	91.1	**65.7**	85.9
STT[12]	100	96.2	95.3	63.0	85.5
Our Model	**100**	**100**	**96.8**	64.1	**87.0**

- The Buffy testset is distributed with a subset of windows detected by a rigid HOG upper-body detector. We compare our results to all previously published work on this subset.
- We obtain the best overall PCP while being orders of magnitude faster than the next-best approaches. Our total pipeline requires 1 second to process an image, while [11, 12] take 5 minutes.

Upper body detection on Buffy Testset

Rigid HOG[8]	Deformable Parts[2]	Our Model
85.1	93.8	**98.6**

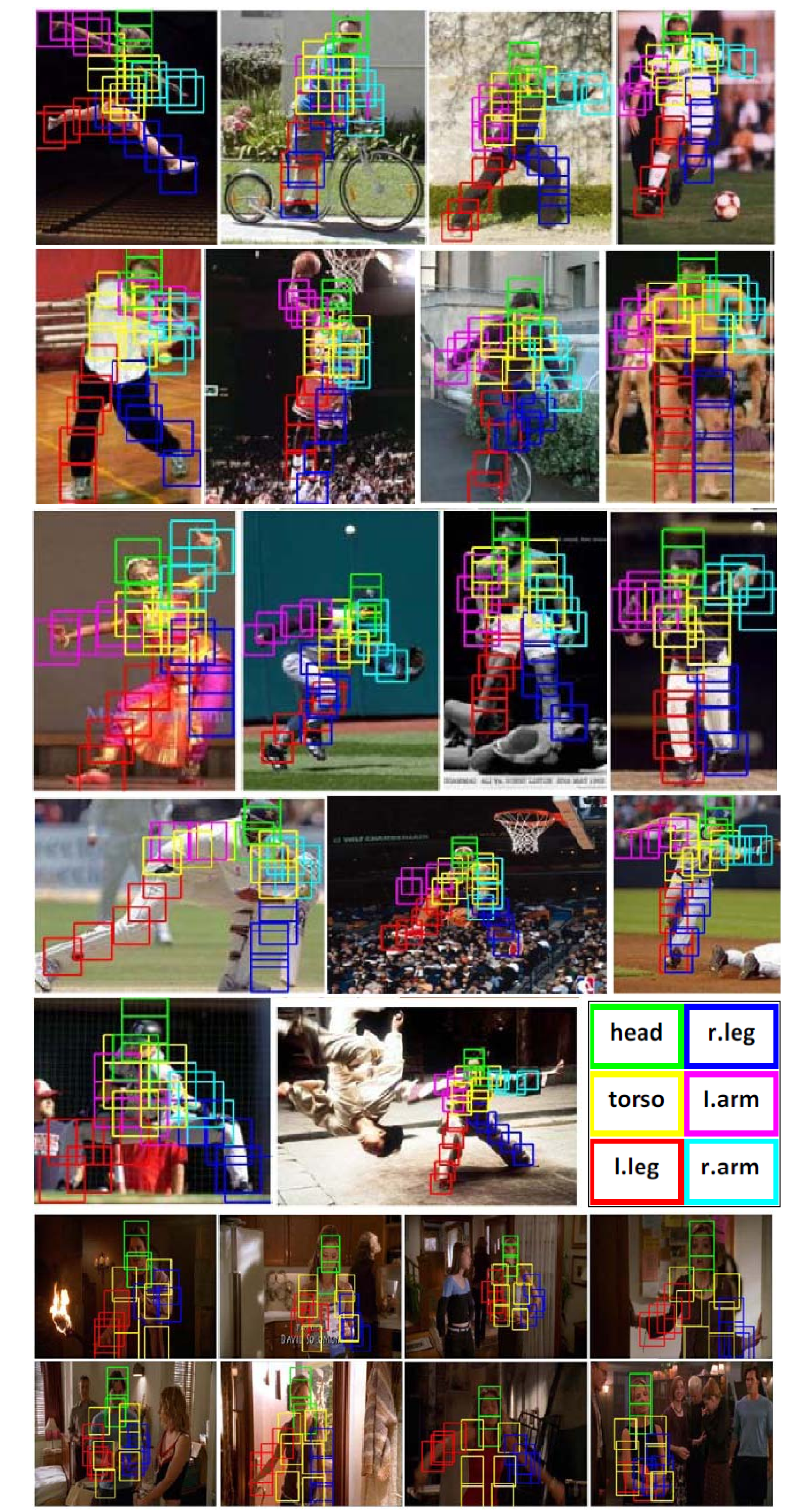
- Our model also serves as an accurate detector. We obtain significantly better upper-body detection results than past work evaluated on the full testset.
- [2] uses a star-structured model of HOG templates trained with weakly-supervised data. Our results suggest more complex object structure, when learned with supervision, can yeild improved results for detection.

Full Buffy Testset

Method	Torso	Head	U.arm	L.arm	Total
TF[9]					53.0
ARS[4]	77.2	81.3	67.5	35.1	62.6
EFZ[10]	84.0	83.4	70.5	50.9	68.2
SJT[11]	85.1	85.1	77.6	55.9	73.1
STT[12]	85.1	81.9	81.1	53.6	72.8
Our Model	**98.6**	**98.6**	**95.4**	**63.2**	**85.7**

- As pointed out by [9], the subset of Buffy testset contains little pose variation because they are biased to be responses of rigid HOG template.
- The distributed evaluation protocol also allows one to compute performance on the full test videos by multiplying PCP values with the overall detection rate.
- Because our model also serves as a very accurate detector, we obtain significantly better results than past work when evaluated on the full testset.

Good Examples



References

[1] Pictorial structures for object recognition. IJCV (2005).
[2] Object detection with discriminatively trained part based models. PAMI (2010).
[3] Learning to parse images of articulated bodies. NIPS (2007).
[4] Pictorial structures revisited: People detection and articulated pose estimation. Proc. CVPR (2009).
[5] Combining discriminative appearance and segmentation cues for articulated human pose estimation. ICCV Workshops (2010).
[6] Efficient inference with multiple heterogenous part detectors for human pose estimation. ECCV (2010).
[7] Clustered pose and nonlinear appearance models for human pose estimation. BMVC (2010).
[8] Progressive search space reduction for human pose estimation. CVPR (2008).
[9] Improved human parsing with a full relational model. ECCV (2010).
[10] Better appearance models for pictorial structures. Proc. BMVC (2009).
[11] Adaptive pose priors for pictorial structures. CVPR (2010).
[12] Cascaded models for articulated pose estimation. ECCV (2010).